Article title: Making interaction with virtual reality accessible: rendering and guiding methods for subtitles

Authors: Belén Agulló (Universitat Autònoma de Barcelona), Mario Montagud (i2Cat Foundation & University of Valencia), Isaac Fraile (i2Cat Foundation)

Corresponding author: Belén Agulló | MRA 020, Campus Universitat Autònoma de Barcelona - Bellaterra, 08193 - Spain | +34 93 586 8108 | belen.agullo@uab.cat

Short title: Subtitles in virtual reality

Pages: 34

Tables: 0

Figures: 11

Title: Making interaction with virtual reality accessible: rendering and guiding methods for subtitles

Abstract: Accessibility in immersive media is a relevant research topic, still in its infancy. This article explores the appropriateness of two rendering modes (fixed-positioned and always-visible) and two guiding methods (arrows and auto-positioning) for subtitles in 360º video. All considered conditions have been implemented and integrated in an end-to-end platform (from production to consumption) for their validation and evaluation. A pilot study with end-users has been prepared and conducted with the goals of determining the preferred options by users, the options that results in a higher presence, and of gathering extra valuable feedback from end-users. The obtained results reflect that, for the considered 360º content types, always-visible subtitles are more preferred by viewers and received better results in the presence questionnaire than the fixed-positioned subtitles. Regarding guiding methods, participants preferred arrows over auto-positioning, because arrows were considered more intuitive and easy to follow and reported better results in the presence questionnaire.

Keywords: virtual reality, 360º videos, human computer interaction, subtitles, accessibility

1. Introduction

There is a growing interest in Virtual Reality (VR) and the possibilities to develop immersive contents, such as 360º videos. Viewers can watch 360º clips with head-mounted displays (HMD) or directly from a flat screen on a smartphone or a computer. In these videos, the viewers have the freedom to look around in the synthetic world and explore the virtual scenarios that are shown to them. YouTube, Facebook, Jaunt VR, The New York Times VR are some of the companies that are developing immersive experiences for their audience via online platforms. According to a report issued by the European Broadcasting Union (EBU, 2017), 49% of its members are exploring and devoting efforts developing immersive content. EBU members believe that immersive content presents a clear potential for broadcasters and content creators, because it offers the opportunity to provide more interactive and engaging storytelling. For content creators and filmmakers, one of the main challenges when developing immersive content is the lack of control over the main focus of the video. Therefore, intelligent and effective strategies to present the contents, attract and keep audience's attention and assist users need to be explored and adopted. Nonetheless, interactive and immersive content creation development are still at an early stage and research on these topics is ongoing (Dooley, 2017; Mateer, 2017; Rothe et al., 2017; Sheikh et al., 2017).

Apart from open challenges in terms of high-resolution contents, interaction and storytelling formats for immersive media, a key issue needs to be taken into account: accessibility. It is not acceptable to consider accessibility as an afterthought, but it instead must be addressed in the specification and deployment of end-to-end immersive systems and services. Such an objective involves overcoming existing limitations in current technologies and systems to enable truly inclusive, immersive and personalized experiences, adapted to the needs and/or preferences of the users.

Although immersive technologies and contents are on the rise, research studies on, and thus solutions for, accessibility in immersive media are limited so far. This hinders the interaction of part of the population with VR experiences. Proper technological solutions, interfaces and recommendations need to be sought in order to ensure a proper narrative, interpretation of contents and usability, regardless of the capacities of the users, their age, language, and/or other specific impairments. This will contribute to a global e-inclusion, offering equal opportunities of access to the whole consumers' spectrum, while ensuring compliance with regulatory guidelines (e.g., Human Rights Obligations).

Many efforts must be devoted on providing efficient solutions and meaningful insights to, among others, the following research questions on this field:

- What are the requirements to enable truly accessible and inclusive immersive services?
- How current (immersive) technologies and systems can be augmented to seamlessly integrate and support accessibility services?
- What kind of assistive technologies can contribute to a better accessibility in immersive media?
- Which presentation modes for accessibility contents are better suited for specific content types?
- What personalization features should be provided to meet users' needs and preferences?
- What benefits are provided (e.g., in terms of usability, content comprehension, level of immersion and engagement, etc.)? How to properly evaluate and define them?

By comparing with traditional audiovisual contents, the integration of access services (i.e., subtitles, sign language interpreting and audio description) faces two main challenges. First, there is more information to process and users can become overwhelmed. Second, the presentation is no longer purely time-based, but it involves a spatial dimension, determined

either by both the user's Field of View (FoV) and by the direction where the main actions are taking place.

This also applies to subtitles, which is one of the most mainstreamed access services, being provided by major TV channels, like BBC (Armstrong et al., 2015), and Video on Demand platforms, like Netflix, HBO or Amazon Video. Subtitles are not only beneficial for viewers with hearing impairments, but also for users with visual impairments if their presentation format can be customized, for non-native speakers, to support the comprehension of contents, and in noisy / public environments where the audio cannot be listened or cannot be turned on. Beyond contributing to overcome audiovisual barriers, the applicability of subtitles enters the realm of other forms of social integration, can have an impact on education and on therapy, and can contribute to increase the engagement and Quality of Experience (Montagud et al., forthcoming).

This article focuses on two essential issues in this research area: how to integrate and present subtitles in 360º videos without breaking immersion and how to guide the users for a more effective and a non-intrusive interaction and storytelling comprehension. The research tasks are being devoted after having conducted user-centric activities to gather requirements from which the proposed solutions have been derived (Agulló and Matamala, forthcoming; Agulló et al., forthcoming). Two strategies are proposed and assessed for subtitle rendering modes: 1) always-visible —the subtitles are anchored to the FoV, always in the same position; and 2) fixed-positioned —the subtitles are rendered in three fixed positions, evenly spaced every 120º around the 360º sphere. Likewise, two strategies are proposed and assessed for guiding methods, when making use of the always-visible rendering mode: 1) arrows —a visual element (arrow) is integrated in the subtitle to indicate the viewers where they need to look at to find the speaker; and 2) auto-positioning —an intelligent strategy that adaptively adjusts the FoV to match the position of the targeted speaker(s) / main action(s). Both strategies have been developed and tested in a pilot study. Their integration in an end-to-end platform, paying special

attention to the player side, the followed evaluation methodology, and the obtained results regarding the impact on immersion and the participants' preferences are reported in this article.

The rest of the article has been structured as follows. In Section 2, state-of-the-art work are reviewed. In Section 3, an overview of the developed end-to-end platform for integration of accessibility services in immersive media is provided. This platform has served as the framework for integrating the presented contributions and conducting the pilot study. Next, the evaluation setup, methodology and obtained results are reported. Finally, the results and their scope are discussed, and some ideas for future work are provided.

2. Related work

VR as a form of entertainment, especially in the form of 360º contents or cinematic VR (Mateer, 2017), has attracted the interest of the research community and industry from different perspectives. There are several studies on narrative in VR (Aylett & Louchart, 2003; Dooley, 2017; Gödde et al., 2018) focused on better understanding the complexities of this new medium. Other studies are tackling the specific topic of focus and attracting attention in VR (Mateer, 2017; Rothe et al., 2017; Sheikh et al., 2017). In addition, some researchers have carried out studies on the impact of cinematic VR on immersion (De la Peña et al., 2010; Cummings & Bailenson, 2016; Jones, 2017) and engagement (Wang, Gu & Suh, 2018).

However, research on subtitling in immersive contents is limited. There are few exceptions. The study carried out by the BBC (Brown et al., 2018) was the first considering this topic and proposing some solutions. The main challenges identified by BBC research team when developing subtitles for immersive content are (Brown et al., 2018):

- there is no area in the scene that is guaranteed to be visible to the viewer, so it is not possible to control what will appear behind the subtitle;
- immersion is important in this medium, so subtitles should not disrupt that experience;

- if subtitles are located outside of the FoV, then the effort to find them should be minimum;
- and including subtitles should not worsen the VR sickness effect.

Based on that and on previous studies, BBC developed four solutions for subtitle rendering:

a) Evenly spaced: subtitles are equally spaced with a separation of 120º in a fixed position below the eye line;
b) Follow head immediately: subtitles follow the viewer as he/she looks around, displayed always in front of the him/her;
c) Follow with lag: subtitles appear directly in front of the viewer, and they remain there until the viewers look somewhere else; then, the subtitles rotate smoothly to the new position in front of the viewer; and
d) Appear in front, then fixed: subtitles appearing in front of users, and then fixed until they disappear (in this case, the subtitles do not follow the viewer if they look around).

They tested the different rendering modes with several clips (different durations: from 1 to 2 and a half minutes) and they concluded that “Follow head immediately” (in our study, always-visible) was the most suitable, according to users’ feedback (Brown et al., 2018). The reasons were that the implementation was easy to understand and subtitles easy to locate. Also, it gave viewers the freedom to navigate the 360º content without missing the subtitles. However, users complained about the blocking effect, i.e. subtitles were blocking important parts of the image and were considered obstructive.

Following the above results, Rothe et al. (2018) also carried out a user study comparing two rendering modes: always-visible (which was called static subtitles in their study) and fixed-positioned (which was called dynamic subtitles in their study). They also tested speaker identification methods based on each mode and included name tags for each speaker. Participants did not state a clear preference for any of the methods. However, the results

regarding key aspects of the VR experience (presence, sickness and workload) favored the fixed-positioned subtitles (Rothe et al., 2018). In both studies, there is no clear solution and further testing is encouraged. To shed some light on these open issues, we decided to test these two methods again with longer and different contents. We also decided to measure presence with the igroup presence questionnaire (IPQ) (http://www.igroup.org/pq/ipq/index.php) to compare the impact of each method on viewers' immersion, if any. As explained in the methodology, IPQ is suitable for this type of contents and the measurements provided are accurate for our purpose. In other studies, Presence Questionnaire (Witmer & Singer, 1998) was used (Rothe et al., 2018). This questionnaire includes a range of questions about interaction and control in the virtual world, which is not suitable for a 360º video with a passive observer. In the BBC study, only one question was asked regarding immersion "I felt immersed in the scene, like I was there" (Brown et al., 2018). Therefore, the results are limited, and our approach contributes to provide more information about the impact of the different subtitle rendering and presentation modes on presence.

To the best of our knowledge, no guiding methods for subtitles have been tested so far. This feature is especially important if subtitles are aimed at viewers with hearing impairments, or when the audio cannot be listened (e.g. noisy or public environments). When the audio cue is missing, support on how to locate the speakers and the main actions in the 360º scene is necessary. There are some studies, though, that tested different guiding methods for assisting focus in 360º videos, which are somehow related with, or have an impact on, guiding methods for subtitling. In particular, some studies have tested different types of transitions and their impact on immersion and motion sickness. The preliminary results from the study by Men et al. (2017) concluded that the transition techniques being tested (Simple Cut Transition, Super Fast Transition, Fade Transition and Vortex Transition) do not cause much sickness, contrary to what could be expected. The study carried out by Moghadam and Ragan (2017) concluded that each

tested transition technique (Teleportation —involves an instant change in current FoV or rotation that is not perceived by the viewer; Animated Interpolation —smooth FoV transition from one state to another, which can be seen by the viewer; and Pulsed Interpolation —the pulsed view is faded in and out to different intermediate points from one state to another) had a different impact on the levels of presence and different techniques should be used depending on the desired effect. Lin et al. (2017) conducted an extensive study comparing two techniques to guide users to the focus of the 360º video: Auto Pilot —a method that takes the viewer directly to the intended target, and Visual Guidance —a visual indicator that signals where the users should direct their view. The goal of this study was to establish which technique was better suited for the viewing experience when focus assistance is necessary. They concluded that both guiding methods were preferred by participants than no guiding method at all for focus assistance. They also argued that the specific content scenario and environment have an impact on which techniques are preferred by users.

In our study, the first goal was to gather participants' feedback (preferences and impact on presence) about two subtitle rendering modes (always-visible and fixed-positioned). In this regard, we have tried to overcome the limitations pointed out in previous studies (Brown et al., 2018), by using longer contents (clips are longer than two minutes) and different genres. We decided to use travel documentaries where the main goal was to have a look at the landscapes and listen to the narrator (voice-over), and thus become a suitable genre to test these research aspects. The following reasons support the decision on choosing this content genre: participants would have freedom to look around without a main focus; no narrative complexities were introduced to avoid confusion; and there were no speakers on screen. Because there were no speakers, we could isolate the variables (rendering modes), without introducing any guiding methods, tested in the second part of the experiment. The second goal of the study was to gather participants' feedback (preferences and presence) about two guiding methods (arrow and

auto-positioning) to determine which system is preferred by users and to find out which method results in a higher immersion.

3. End-to-End Platform for Immersive Accessibility

This research work has been conducted within the umbrella of the EU H2020 Immersive Accessibility (ImAc) project (October 2017 - March 2020, http://www.imac-project.eu/). By following a user-centric methodology, ImAc is exploring how accessibility services (subtitling, audio description, and sign language interpreting) can be efficiently integrated within immersive media (360º video, spatial audio, and VR contents), while enabling different interaction modalities and personalization features. To achieve the targeted goals, ImAc is developing an end-to-end platform comprised of different parts where production, edition, management, preparation, delivery and consumption of (immersive and accessibility) contents take place. Figure 1 provides a high-level overview of the logical layers or main parts of the ImAc platform, which adhere to current-day media broadcast and delivery workflows. In this figure, green color is used to identify components being developed within the ImAc project, orange color is used to identify components that are relevant for ImAc, but that have been developed in other related projects, and white color is used for components that exist in typical broadcast workflows, but that are either not part of or not essential for ImAc. Next, an overview of each one of the platform parts is provided to better understand the context - and potential impact - of this work.

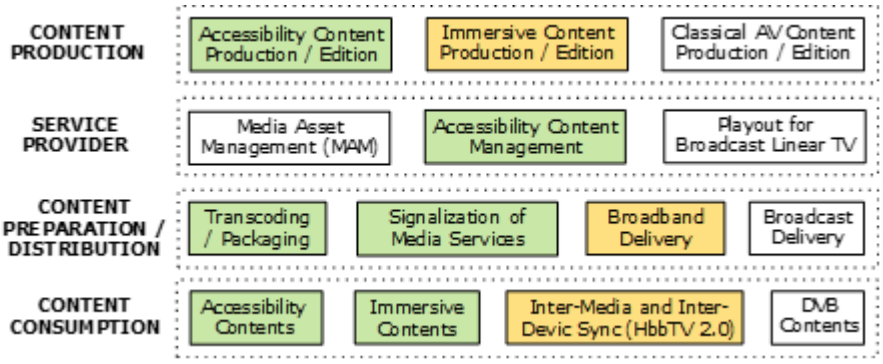

Figure 1. Main layers or parts of the end-to-end platform

3.1. Content Production

The Content Production part of the platform includes a set of (web-based) tools for the production and edition of access services (including subtitles, audio description and sign language interpreting), and their integration with immersive media contents. The subtitles production / edition tool enables the creation of subtitles for 360º videos. Unlike existing editors that mainly allow the production of subtitles frames with specific timing attributes (i.e. start and end times), the ImAc editor provides a set of additional features targeted at contributing to a better accessibility (and engagement) [Footnote 1]:

- It allows setting different styling effects (e.g. colors, font) for different speakers.

- It allows indicating spatial attributes to set the region of the 360º area to which the subtitle frames refer. The spatial information consists of the latitude and longitude angles (although only the latitude ones are considered in this work). This is relevant, as the associated action(s) / speaker(s) can be placed in different parts of the 360º area and can even dynamically move. However, it is possible to indicate that no spatial information is linked to specific subtitle frames (e.g. for off-camera commentary).

- It allows specifying two options for subtitles rendering (in flat format). The first option consists of using the 360º sphere as the rendering reference (see Figure 2). This is called fixed-positioned, as subtitles are attached (i.e. statically placed) in a fixed region of the video sphere. Using this mode, subtitles will not be visible if the user's FoV is outside the subtitles region. To overcome this, subtitles can be presented evenly spaced every 120º in the 360º sphere, ensuring at least one of them will be visible at any time, regardless of the current FoV. The second option consists of using the current FoV as a rendering reference (see Figure 3). This is called always-visible, as subtitles are attached to the camera, and thus positioned in the center of the FoV at any moment, regardless of where the user is looking at.

Figure 2. Fixed-positioned: subtitles attached to the video sphere

Figure 3. Always-visible: subtitles attached to the camera or FoV

- When using always-visible subtitles, it allows specifying different guiding methods to assist the users in finding the action(s) / speaker(s) associated to the subtitles in the 360º area. A first option consists of adding arrows to the left / right of the subtitle frames, indicating the direction

towards the associated audiovisual elements are in the 360º area (see Figure 4). When this position is inside the user's FoV, the arrows are hidden. A second option consists of automatically adjusting the FoV based on the position of the associated action(s) / speaker(s). This auto-positioning mechanism is applied to every subtitle frame with spatial information, if explicitly indicated in the editor.

All these rendering and presentation features are signalized as metadata extensions to the IMSC (Internet Media Subtitles and Captions) subtitles format, being used in ImAc. IMSC is a subset of the TTML (Timed Text Markup Language) for distribution of subtitles, which is drawing the attention of, and being adopted by, many standardization bodies. Finally, the editor also allows importing and converting existing traditional subtitles files, by adding to them the required metadata for their adequate presentation in 360º videos.

Figure 4. Subtitles with arrows as a guiding mechanism

3.2. Service Provider

This part of the platform includes components for Media Asset Management (MAM), linking of additional contents to main TV programs, and scheduling playout. In the context of ImAc, it additionally includes the Accessibility Content Manager (ACM), which is the component where the immersive contents are uploaded, the creation of accessibility contents is managed, and the preparation of contents for their delivery is triggered.

3.3. Content Preparation & Distribution

This part of the platform includes components for preparing the uploaded and produced contents for their appropriate distribution via various technologies. These components are in charge of encoding the contents in multiple qualities (to adapt to the target consumption devices and available bandwidth), segmenting the contents for an efficient quality adaption and re-transmission (e.g. in case of packet loss), signaling their availability, and describing them. The project focuses on the delivery of the contents via broadband Content Delivery Networks (CDNs), by making use of Dynamic Adaptive Streaming over HTTP (DASH) as the media delivery technology. However, it is also envisioned to make use of DASH in coordination of Digital Video Broadcasting (DVB) services, as supported by the worldwide adopted Hybrid Broadcast Broadband TV (HbbTV) standard. This will enable augmenting traditional TV services with more interactive and personalized multi-screen experiences, enriching the traditional TV contents with extra immersive and accessibility contents presented on companion devices, like smartphones or even HMDs.

In this context, ImAc is exploring the specification of standard-compliant extensions to media formats and technologies (e.g. within the framework of Moving Picture Experts Group or MPEG) to accommodate the envisioned immersive accessibility services and features.

3.4. Content Consumption

The ImAc player is a core component of the ImAc platform, as it is the interface through which end-users will consume the available immersive and accessibility contents in an interactive and personalized manner. The design and implementation of the player faces many challenges due to a number of facts, such as the nature and combination of media contents to be consumed, the heterogeneity in terms of access networks and consumer devices to be employed, the diverse needs and/or preferences of the target end-users, etc.

The player has been developed by exclusively relying on standard(-complaint) web-based technologies and components. This will guarantee cross-network, cross-platform and cross-browser support, and eliminate the need for any installation or software updates. The use of web-based components also facilitates the embedding of the player within the web services of broadcasters and/or service providers, ensuring interoperability and scalability.

Figure 5 illustrates the main layers and modules and libraries that make up the player, together with the relationships and interactions between them. All these components are mainly targeted at enabling the presentation of contents, to enable different interaction features, and to dynamically set the available personalization options.

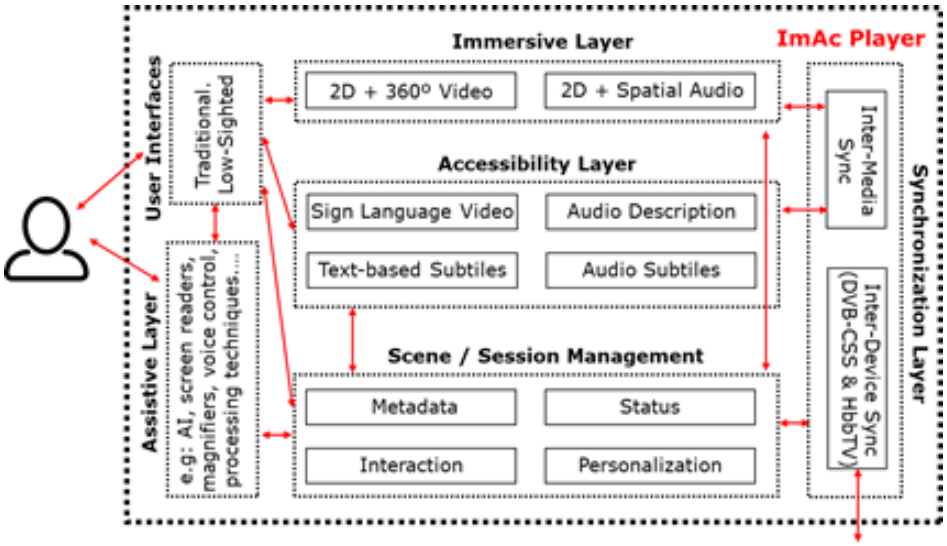

Figure 5. Layers and Modules making up the immersive accessibility (ImAc) player

Three main layers are in charge of the presentation of contents in the player. These include:

- The Immersive Layer: it is responsible for the presentation of both traditional and immersive audiovisual formats. For immersive media, it includes 360º videos and spatial audio (Ambisonics).
- The Accessibility Layer: it is responsible for the presentation of accessibility contents considered in the project, namely: audio and text subtitles; audio description; and sign language video.
- The Assistive Layer: it includes relevant features to assist the users for a more effective usage of the player. Some examples are: voice control (recognition and synthesis), augmentation / zooming capabilities, Artificial Intelligence (AI) techniques, and media processing techniques to improve the interpretation of contents.

Likewise, the Media Synchronization Layer is in charge of ensuring a synchronized consumption of contents, both within each device (i.e. local inter-media synchronization) and across devices in a multi-screen scenario (i.e. inter-device synchronization).

In addition, two main modules in the ImAc player can be highlighted:

- The User Interface (UI): it is the module through which users enable the presentation of contents, interact with the player and set the available personalization features. Indeed, two UIs have been designed and implemented: 1) a traditional UI but adapted to 360º contents (see Figure 6); and 2) an enhanced-accessibility (aka low-sighted) UI, which occupies most part of the screen (see Figure 7).
- The Session Manager: it is the module responsible for interpreting and selecting the list of available assets from the content provider, keeping an updated status about the

contents being presented and the active devices in multi-screen scenarios, and keeping track of the available personalization options together with the current settings.

Figure 6. Screenshot of the Traditional UI of the player

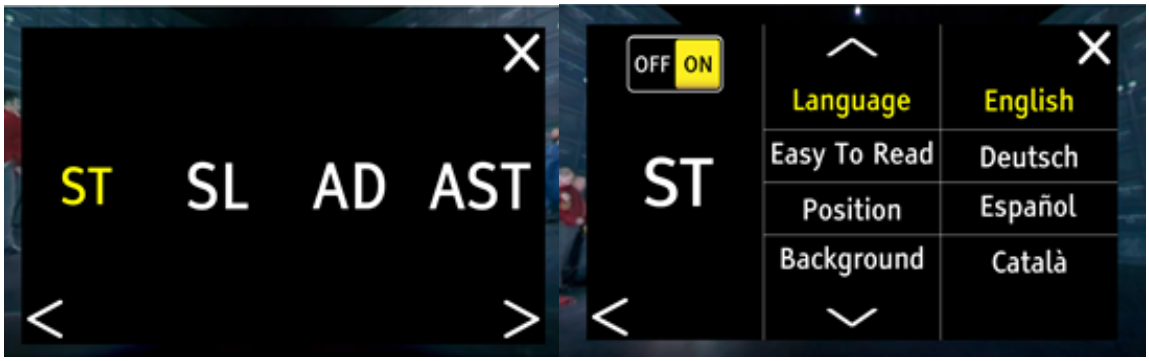

Figure 7. Screenshots of the Enhanced-Accessibility (or Low-Sighted) UI of the player

Content Presentation Modes

The player can be used by any device with a web browser and Internet connection, ranging from connected TVs, smartphones, PCs to VR devices, like HMDs or VR cardboards. Depending on the device's capabilities, the player enables two presentation modes: 1) tablet mode: based on the use of the touch screen, gyroscope sensor, mouse and keyboard for navigation and

interaction; 2) VR mode (for VR-enabled devices): use of HMD buttons, movement trackers and controllers for navigation and interaction.

Playback Control Commands

The player includes the typical playback control and provides visual feedback when clicking on each one of them (see Figure 8). On the one hand, the size of the control is reduced, and its color is changed to yellow (for high contrast) for a short period of time. On the other hand, the current setting / state is shortly displayed.

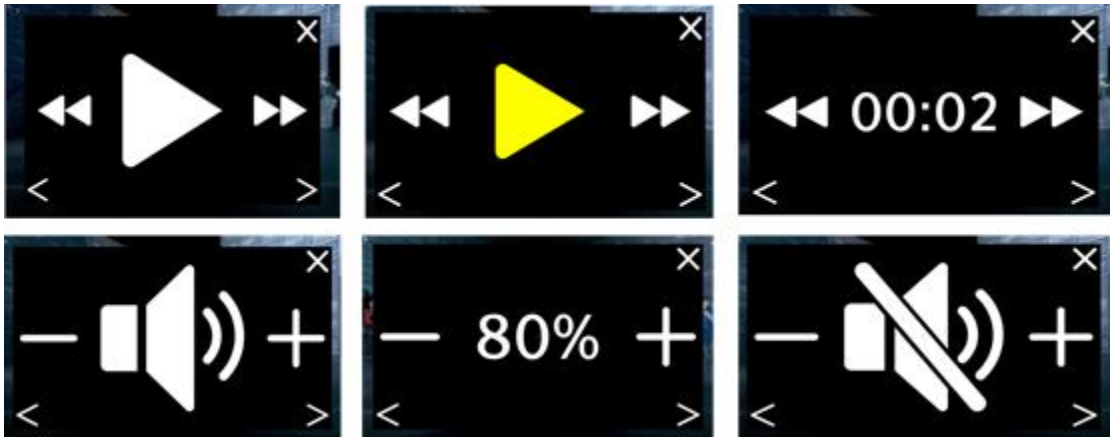

Figure 8. Screenshots for Playback Control and Visual Feedback

Personalized Presentation of Accessibility Services

The player provides support for a personalized presentation of access services, including subtitles, audio description and sign language video (see UIs in Figures 6 and 7). Most interestingly for this work, the player allows dynamically setting the following personalization features for presentation of subtitles:

- Language selection
- Three sizes for the Subtitle font (Large, Medium, Small)

- Position (top and bottom)
- Three sizes for the safe area or the Comfortable FoV where to place graphical elements on the screen. Although the screen size and resolution of the device in use is automatically detected, users can have different preferences regarding this aspect.
- Background (semi-transparent box for the subtitles frame, outline)
- Normal vs easy-to-read subtitles (i.e. more simple and shorter subtitles)
- Guiding methods: 1) None; 2) Arrows indicate where the associated speaker is; and 3) Auto-positioning: the FoV is automatically adjusted based on the location of the speaker. An additional method is available, which consists of using a dynamic radar to indicate where the associated speaker or main action is (see Figure 9). However, this method has not been tested in this work, as pre-tests have indicated a preference towards the arrows.

Figure 9. Use of a Dynamic Radar as a guiding method

Apart from the user-level personalization features, the rendering mode for subtitles and different styling effects (color, font, etc.) for each speaker can be set during the production / edition phase (at the Content Production part).

Interaction Modalities and Assistive Technologies.

The player supports different interaction modes and modalities. When using a PC / laptop as the consumption device, interaction with the player can be done via the mouse and/or keyboard. When using a tablet or smartphone, interaction with the player can be done via the touch screen, although it is also possible to navigate around the 360º area by using the gyroscope sensor. When using a VR-enabled device, the head movement sensors can be used to navigate around the 360º area and to move the cursor. The controllers of the VR device can also be used to move the cursor, while their touchpad and physical buttons can be used to select and/or navigate between menus of the UI.

In addition, voice control is also available as an interaction modality and assistive technology. It is provided by making of the World Wide Web Consortium (W3C) Web Speech Application Program Interface (API). In particular, the *SpeechRecognition* part of this API is used for asynchronous speech recognition, and the *SpeechSynthesis* part for text-to-speech synthesis (i.e. to provide spoken feedback to execution of commands).

Other assistive features being considered include: zoom and enlargement functionalities for visual menus and/or controls; use of screen-readers; and use of media processing techniques for a better accessibility. The use of AI techniques will be key in maximizing the efficiency of such assistive features.

Integration in Multi-Screen Scenarios

Apart from the consumption of immersive and accessibility contents within single devices, the player is prepared for their integration in multi-screen scenarios (Fraile et al., 2018). In such cases, a main screen (e.g. a connected TV) will play out traditional TV contents (plus optionally accessibility contents), and one or multiple companion screens will play out the immersive contents (i.e. 360º videos and spatial audios) in combination with the accessibility contents, as

can be seen in Figure 10. This enables more personalized, accessible and engaging media consumption experiences.

Figure 10. ImAc player integrated in multi-screen scenarios

4. Evaluation

An experiment to test two different aspects of subtitles in 360º content (rendering and guiding methods) was conducted. The goal of this experiment was to clarify which options are preferred by users, as well as which ones result into higher presence. This section describes the selected and created stimuli for conducting the tests, the evaluation scenario and setup, the followed evaluation methodology and presents the obtained results.

4. 1. Evaluation stimuli

An acclimation clip was introduced at the beginning of the test, so that participants could become comfortable with the HMD and the type of content, assuming that most participants did not have an extensive experience with the use of HMD and VR experiences. This was later confirmed by the replies to the demographic questionnaire. All clips included sound (voice over in English), because it was considered that sound is an important part of the immersive

experience and presence was being measured as part of the test. Subtitles were produced in Spanish, a language spoken and understood by all participants.

For the first condition (rendering modes), two videos from the series The Holy Land created by Ryot [Footnote 2] jointly with Jaunt VR were used. Creators gave their permission to include their videos in the study. Specifically, the episode 4 (duration of 4 minutes and 13 seconds) [Footnote 3] and 5 (duration of 4 minutes and 58 seconds) [Footnote 4] were chosen. The clips are travel documentaries depicting Israel and surrounding territories. Different locations and landscapes are featured. In the clips, there is only one speaker and most of the script is voice-over (narrator), except from some scenes where the hostess can be seen. The videos were considered suitable for testing the subtitle rendering modes, because viewers could concentrate on reading subtitles and watching the scenes without the added effort of having to look for the speakers or any other narrative complexities.

For the second condition (guiding methods), the clip An American Classic: Guelaguetza [Footnote 5] also created by Ryot was used. In this case, the video was split into two parts, in order to have two comparable clips. The total duration of the clip is 7 minutes and 58 seconds (first part from 00:00 to 03:46 and second part from 03:46 to 07:16 —credits start). This short documentary narrates the story of a family from Oaxaca that decided to immigrate to Los Angeles and opened a restaurant there. In the video, the two generations of owners (mother and daughter) explain their experiences and what the restaurant and their food mean to them. The clip combines scenes with different locations and a voice-off narration with scenes were Bricia (daughter) and Maria (mother) appear explaining their experiences. This video is suitable for the test, because it includes different people in different locations, which would elicit viewers search for the speakers. Also, locating the speakers in the video does not require a great effort (they are mostly located in the same area, standing or sitting), which was also desirable for the test to

avoid confusion among viewers, especially for those ones not being familiar with any of the guiding methods or with VR technology in some cases.

4.2. Evaluation Setup

The evaluations were conducted in a local scenario with of a PC with an Apache web server (no high computational resources are required) to host the player resources and the media assets (360º video and subtitles), a conventional 802.11b WiFi network and a standalone VR Oculus GO (32GB) as consumption device. The Oculus GO accessed the player via its WiFi connection and by typing the target URL pointing to the server resources. Note that the web server and clients could have been placed in remote locations, and that other types of consumption devices, and other HMDs, could have been used.

The 360º videos were converted into DASH format, being encoded in multiple qualities (with bit rates ranging from 8Mbps to 2Mbps) and segmented in chunks of a duration of 3s. This allows an efficient quality switching adaptation, based on the network and consumption devices conditions. The subtitle files were delivered independently to the video segments, but they were signalized as part of the video metadata files. An overview of the evaluation scenario and setup can be seen in Figure 11.

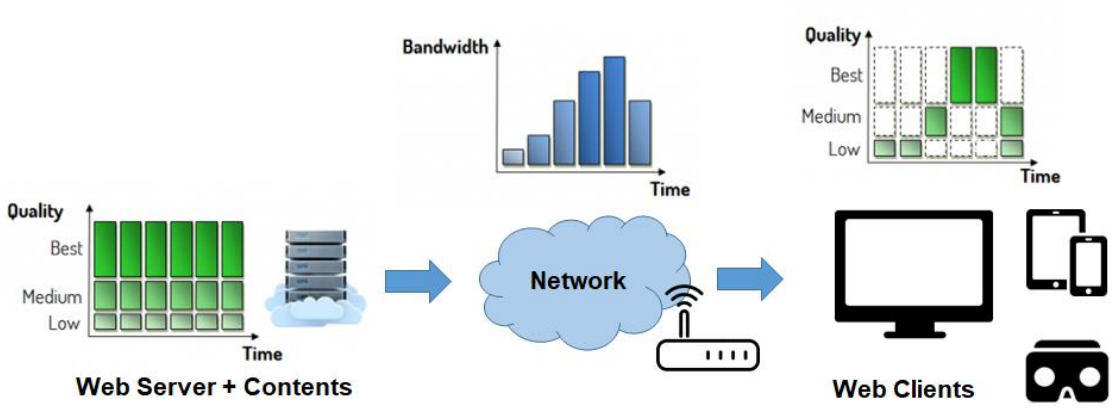

Figure 11. Overview of the evaluation scenario and setup

4. 3. Evaluation methodology

A within-subject design was used to test the different subtitle presentation conditions. Each participant watched four clips (plus the acclimation video), being each of them presented with a different variable (fixed-positioned, always-visible, arrow and auto-positioning). The four clips and four conditions were randomized and counterbalanced using a Latin square, to avoid the order of presentation affecting the results. The clip An American Classic: Guelaguetza was always shown in chronological order, otherwise the participants would have not been able to understand the story.

The experiment was organized in one session divided into two parts: Part 1 - Rendering modes, and Part 2 - Guiding methods. The experiment was focused on assessing users' preferences and presence. One of the main goals of immersive content, such as 360º videos, is to create an immersive experience. Therefore, it was paramount to design subtitles that would enhance the experience making it more accessible rather than disrupting it. Likewise, an additional goal of the test was to gather feedback from the users. This will allow deriving potential requirements for improving the provided functionalities or even incorporating additional ones, thus following the user-centric methodology being used in ImAc. To gather this feedback, questionnaires were used.

For presence, a translation into Spanish of the IPQ questionnaire was used. After a review of different presence questionnaires, such as Slater-Usoh-Steed Presence Questionnaire (Slater & Usoh, 1993), Presence Questionnaire (Witmer & Singer, 1998) or ICT-SOPI (Lessiter et al., 2001), IPQ was chosen for different reasons. First, it includes questions from different questionnaires and it specifically differentiates between presence, spatial presence, involvement and realness. The questionnaire has been validated in different virtual environments (users of VR or CAVE-like systems, desktop VR, players of 3D games, etc.). Also, unlike other questionnaires, such as the Presence Questionnaire by Witmer and Singer (1998), the

questions in IPQ do not involve interaction with the virtual world. This was important, because the 360º clips that were chosen for the test are not interactive.

For preferences, an ad-hoc questionnaire in Spanish for this test was created for each part (rendering and guiding methods). The questionnaires included closed questions to assess which system users preferred and questions related to subtitles' blocking or distracting effects. Also, open questions were used to gather feedback about the reasons to choose one method over the other, and 7-point Likert-scale questions were added to determine how easy was to find or read subtitles, as well as to find the speaker in the video.

After watching each clip, participants were asked to reply the IPQ questionnaire, so that the level of presence could be later compared between the two options. The impact of the different subtitle strategies on presence, if any, could then be measured and reported. After each part, participants were also asked to reply preference questionnaires, so that they could report on their experience with both options for rendering and guiding methods.

4. 4. Participants

Eight participants took part in the test (three female and five male), with ages ranging from 26 to 59 (standard deviation of 13.18). Two participants were deaf. Our aim was to include different profiles of subtitle users to gather relevant feedback in this preliminary study. To that end, users from different ages and hearing abilities were included. As explained before, subtitles are not only beneficial for deaf audience, but also for non-native speakers. This is due to the fact of the wide applicability of subtitles, as discussed in Section 1.

4. 5. Evaluation results

The results from the different questionnaires are reported in this subsection.

Demographic information

Some more demographic information about participants was gathered. Five participants had university education, two had professional training and one had primary education. Two participants were familiar with VR contents (one participant stated to use VR once a week, and another participant, once a month). Five participants were interested in VR contents, and three were neutral. Three participants owned VR equipment: one had a cardboard, another had a PlayStation VR and the last one had a Google Cardboard and a PlayStation VR. Two participants claimed that they never use subtitles, four claimed that they use them sometimes (depending on the content, the language and the context —noisy room, other people watching the content at the same time, etc.) and two of them always used them. When asked about the reasons to use subtitles, one participant said to learn languages, four said that they used them because subtitles helped them to understand, one participant claimed that subtitles are the only way to access the dialogues, and two said that they never use subtitles.

IPQ and preferences

Participants' self-assessed experiences were analyzed based on two types of questionnaires: IPQ to measure and compare the levels of presence, and ad-hoc questionnaires to gather feedback about participants' preferences regarding the considered subtitle presentation modes. The results for the IPQ test aimed at detecting differences in levels of presence between always-visible and fixed-positioned subtitles, and between subtitles with arrows and auto-positioning, and the existence of significant differences between the tested conditions has been analyzed by using a Wilcoxon Signed Rank test with a threshold value of .05. The IPQ is divided in four main blocks: presence, spatial presence, involvement and realness, and the results are reported below based on that classification.

Always-visible vs fixed-positioned

All participants preferred the always-visible subtitle rendering mode. The main reasons for having chosen this option according to the participants is that with the always-visible subtitles they had more freedom to look around without missing the subtitle content and the video scenes. When asked about how easy was to find the always-visible subtitles based on a 7-point Likert scale, six participants (75%) replied 7, one (12.5%) replied 6, and one (12.5%) replied 5. When asked the same question about fixed-positioned subtitles, three participants (37.5%) replied 2, two (25%) replied 3, two (25%) replied 4, and one (12.5%) replied 5. Then, according to these results, always-visible subtitles (mean=5.78) were considered easier to find than fixed-positioned subtitles (mean=3.12). When asked about how easy was to read always-visible subtitles based on a 7-point Likert scale, three participants (37.5%) replied 6, two (25%) replied 2, one (12.5%) replied 7, one (12.5%) replied 5, and one (12.5%) replied 3. When asked the same question about fixed-positioned subtitles, two (25%) replied 6, two (25%) replied 5, two (25%) replied 2, one (12.5%) replied 7, and one (12.5%) replied 3. Therefore, according to these results, always-visible subtitles (mean=4.62) were considered slightly easier to read than fixed-positioned subtitles (mean=4.5). When participants were asked whether subtitles were obstructing important parts of the image, five participants (62.5%) replied “no” and three (37.5%) replied “yes” for always-visible subtitles, and seven participants (87.5%) replied “no” and one (12.5%) replied “yes” for fixed-positioned subtitles.

The comparison of results from IPQ between the always-visible and fixed-positioned are as follows. For the presence scale, the test indicated that the difference between results is not statistically significant (Z=-1.000, p=.317, ties=7). For the spatial presence scale, the test indicated that the difference between results is not statistically significant (Z=-.594, p=.553, ties=1). For the realness scale, the test indicated that the difference between results is not statistically significant (Z=-.318, p=.750, ties=2). However, for the involvement scale, the test reported that the difference between results is statistically significant (Z=-2.032, p=.042, ties=1).

This means that the fixed-positioned subtitles had a negative impact on the involvement of participants. According to their comments in the open questions, this could be because they felt less free to explore the virtual world and claimed to have missed parts of the subtitles content. Moreover, as reported above, participants found more difficult to find subtitles in this mode. Therefore, this extra effort could have caused a negative impact on involvement.

Arrow vs auto-positioning

Seven participants (87.5%) preferred the arrows over the auto-positioning method. Participants who favored the arrows argued that this system is more intuitive and comfortable. Three participants suggested that the arrow guiding mechanism should also include indications for the vertical axis (up, down), not only for the horizontal one (left, right). The participant who preferred the auto-positioning considered that it was more comfortable, because there was no need to move or look for the speaker. One participant also argued that she would like to have a focus assistance technique not only for speakers, but also for main action in the videos. For example, if a specific event is happening in a part of the video (even if no one is speaking), she considered that it would be useful to have an indicator to avoid getting lost in the video. When asked about how easy was to find the speaker with the arrow guiding mechanism based on a 7-point Likert scale, three participants (37.5%) replied 6, two (25%) replied 7, two (25%) replied 4 and one (12.5%) replied 5. When asked the same question about the auto-positioning, three participants (37.5%) replied 7, three (37.5%) replied 1, one (12.5%) replied 6 and one (12.5%) replied 3. The different results in the latter could be because some participants reported feeling dizzy and disoriented with the auto-positioning system and others did not have the same experience. According to the results, arrows (mean=5.62) were considered more effective to find the speaker than auto-positioning (mean=4.12). When asked whether the guiding mechanisms distracted participants from the story, seven participants (87.5%) replied “no” and one (12.5%)

replied “yes” for the arrows, and five participants (62.5%) replied “yes” and three (37.5%) replied “no” for the auto-positioning.

The comparison of results from IPQ between arrow and auto-positioning are as follows. For the spatial presence scale, the test indicated that the difference between results is not statistically significant (Z=-.531, p=.595, ties=2). For the involvement scale, the test indicated that the difference between results is not statistically significant (Z=-.431, p=.666, ties=2). For the realness scale, the test indicated that the difference between results is not statistically significant (Z=.000, p=1.000, ties=1). However, for the presence scale, the test reported that the difference between results is statistically significant (Z=-2.000, p=.046, ties=4). This means that the auto-positioning method had a negative impact on the presence in the virtual world. According to comments in the open questions, participants were not satisfied with the auto-positioning mainly because they felt dizzy, it broke the immersion, and was confusing.

6. Conclusions and future work

This article has investigated the suitability of different rendering modes and guiding methods for subtitles in 360º videos. The considered options have been integrated in an end-to-end platform being developed in the ImAc project. An overview is such a platform has been provided to better understand the context of this work, and its potential impact. According to the obtained results, it can be concluded that always-visible subtitles are more appropriate than fixed-positioned subtitles. These findings are in line with the ones from the study carried out by BBC (Brown et al., 2018), but we have tried to overcome some limitations of that work (such as the duration of the contents). Even if the contents are longer, always-visible subtitles seem to be the most suitable of the rendering modes explored so far. Moreover, in our case, participants did not complain about the blocking effect of the subtitles, as it happened in the BBC study. This could be due to the fact that we did not use a background box in the subtitles, and therefore, they were less intrusive. As explained in Section 3, the use of a background box or an outline can be

dynamically set in the developed player. Also, the results from the IPQ have shed some light on the potential impact of rendering modes on presence levels reported by participants. Fixed-positioned subtitles might have a negative impact on presence, while always-visible subtitles seemed to be more adequate in that sense.

Regarding the two analyzed guiding methods, it can be concluded that the use of arrows is more intuitive and effective than auto-positioning. Even if previous studies argued that auto-positioning methods are accepted by users (Lin et al., 2017), in our study it can be concluded that auto-positioning can provoke dizziness (as reported by participants), and might have a negative impact on presence, at least for the considered content types.

The scope of this preliminary study was to test several subtitle modes with a limited number of participants. Including diverse profiles was sought to clarify the different needs of subtitle users. The selected contents might have had an impact on preferences and presence results that is not directly related to the different subtitle modes. For the rendering modes options, two travel documentaries were used. In this type of contents, the aim is to look around and, then, it is desirable to have freedom to move. However, if the video features a conversation between two people in a bar, perhaps the fixed-positioned solution would be more accepted. A similar content (two people talking and sitting next to each other) was tested in the study by Rothe et al. (2018) and the results favored fixed-positioned subtitles. Also, some participants argued that the videos were not first-person and, therefore, were less immersive. Others thought that the quality, scales and type of scenes also had a negative impact on immersion. For the guiding methods, we used a content where the speakers were mainly in a fixed position and did not rapidly move. Perhaps, if the content includes speakers moving fast, an improved auto-positioning system could assist viewers keeping the focus of the video. These hypotheses are worth testing in future studies. Likewise, a wider sample of participants, with different profiles, will be considered to test the conditions in the near future.

An interesting future work idea for rendering modes could be comparing the effectiveness of always-visible and fixed-positioned subtitles depending on the type of content (static scenes vs action-based scenes), by analyzing whether the type of content has a direct impact on the viewers' preferences and levels of presence. Combining the two rendering modes in a content with different types of scenes (static and action-based) and measure the reaction of participants is also an option worth exploring. Regarding guiding methods, auto-positioning strategies could be refined to reduce the VR sickness effect and test it with other types of contents (action-based). Likewise, the use of a dynamic and intuitive radar (as introduced in Section 3) could be explored in future tests. Finally, the feedback from the participants will be taken into account to explore the suitability of refining the adopted solutions and/or adopting extra alternatives (for example, including guiding methods not only for speakers but also for main actions).

Acknowledgements

Authors would like to thank ImAc partners, for their technical contributions in preparing the end-to-end platform, from production to consumption of immersive and accessibility contents; Irene Tor-Carroggio for her help analyzing the results in this research; Ryot for the contents; and all users for participating in this research.

This work has been funded by European Union's Horizon 2020 program, under agreement nº 761974 (ImAc project).

Belén Agulló is member 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). This article is part of Belén Agulló's PhD in Translation and Intercultural Studies at the Department of Translation, Interpreting and East Asian Studies (Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental) of the Universitat Autònoma de Barcelona.

Author biographies

Belén Agulló is a predoctoral researcher in the Department of Translation, Interpreting and Eastern Asian Studies at the Universitat Autònoma de Barcelona, where she works on the Horizon 2020-funded project Immersive Accessibility (ImAc). Her PhD focus is subtitling for the deaf and hard-of-hearing in immersive media. Her research interests include game localization, audiovisual translation and media accessibility.

Dr. **Mario Montagud** is a Senior Researcher at i2CAT Foundation (Barcelona, Spain) and a Part-Time Professor at the University of Valencia (Spain). He received a PhD degree in Telecommunications (Cum Laude Distinction) at the Polytechnic University of Valencia (UPV, Spain) in 2015. His topics of interest include Computer Networks, Interactive and Immersive Media, Synchronization and QoE (Quality of Experience). Mario is (co-) author of over 70 scientific and teaching publications and has contributed to standardization within the IETF (Internet Engineering Task Force). He is member of the Editorial Board of international journals and has been member of the Organization Committee of the many international workshops and conferences. He is also lead editor of “MediaSync: Handbook on Multimedia Synchronization” (Springer, 2018) and Communication Ambassador of ACM SIGCHI (Special Interest Group in Human-Computer Interaction). He is currently involved in three EU H2020 projects, being WP leader in two of them: VR-Together, ImAc and 5G-Picture.

Mr. **Isaac Fraile** is a network engineer by the Polytechnic University of Catalonia (UPC, Spain) and is currently studying a MSc Degree in Multimedia Applications at the Universitat Oberta de Catalunya (UOC, Spain). He works as a Project Engineer in the Media Unit of i2CAT Foundation. His work is focused on the design and implementation of immersive media platforms, paying special attention to delivery and synchronization techniques, and to web-based components. Previously, he has worked for a year and a half for the CCMA, the public

regional broadcaster in Catalonia. He has participated in three EU projects: TV-Ring, ImmersiaTV and ImAc.

Footnotes

Footnote 1: Further features are planned to be incorporated in the near future, such as the use of visual icons augmenting the textual information.

Footnote 2: https://www.jauntvr.com/lobby/ryot

Footnote 3: https://www.jauntvr.com/title/b4f85188a2

Footnote 4: https://www.jauntvr.com/title/fb1051a266

Footnote 5: https://www.youtube.com/watch?v=zneKYGQgabk

Figures

References

Agulló, B. & Matamala, A. (forthcoming). The challenge of subtitling for the deaf and hard-of-hearing in immersive environments: results from a focus group, *The Journal of Specialised Translation 32*.

Agulló, B., Matamala, A., & Orero, P. (forthcoming). From disabilities to capabilities: testing subtitles in immersive environments with end users. *HIKMA 18*.

Armstrong, M., Brown, A., Crabb, M., Hughes, C., & Sandford, J. (2015). Understanding the Diverse Needs of Subtitle Users in a Rapidly Evolving Media Landscape. *IBC 2015*, Amsterdam (The Netherlands), September 2015.

Aylett, R., & Louchart, S. (2003). Towards a narrative theory of virtual reality. *Virtual Reality 7(1)*, 2-9.

Brown, A., Turner, J., Patterson, J., Schmitz, A., Armstrong, M., & Glancy, M. (2018). *Exploring Subtitle Behaviour for 360° Video*. *White Paper WHP 330*. BBC.

Cummings, James J., & Bailenson, Jeremy N. (2016). How Immersive Is Enough? A Meta-Analysis of the Effect of Immersive Technology on User Presence. *Media Psychology 19(2)*, 272-309.

De la Peña, N., Weil, P., Llobera, J., Giannopoulos, E., Pomés, A., Spanlang, B., Friedman, D., Sanchez-Vives, M. V., & Slater, M. (2010). Immersive Journalism: Immersive Virtual Reality for the First-Person Experience of News. *Presence: Teleoperators and Virtual Environments 19(4)*, 291-301.

Dooley, K. (2017). Storytelling with virtual reality in 360-degrees: a new screen grammar. *Studies in Australasian Cinema 11(3)*, 161-171.

European Broadcasting Union (EBU) (2017). *Virtual Reality: How are public broadcasters using it?* Retrieved from: https://www.ebu.ch/publications/virtual-reality-how-are-public-broadcasters-using-it (consulted on 28.11.2018).

Fraile, I., Gómez, D., Núñez, J. A., Montagud, M., & Fernández, S. (2018). Personalized and Immersive Presentation of Video, Audio and Subtitles in 360º Environments: An Opera Use Case. *ACM TVX 2018*, Seoul (South Korea), June.

Gödde M., Gabler F., Siegmund D., & Braun, A. (2018). Cinematic Narration in VR – Rethinking Film Conventions for 360 Degrees. In *Virtual, Augmented and Mixed Reality: Applications in Health, Cultural Heritage, and Industry* (Chen, J., & Fragomeni, G., Eds), VAMR 2018. Lecture Notes in Computer Science, Vol. 10910. Cham: Springer.

Jones, S. (2017). Disrupting the narrative: immersive journalism in virtual reality. *Journal of Media Practice 18(2–3)*, 171–185.

Lessiter, J., Freeman, J., Keogh, E., & Davidoff, J. (2001). A Cross-media Presence Questionnaire: The ITC-Sense of Presence Inventory. *Presence: Teleoperators, and Virtual Environments 10(3)*, 282–297.

Lin, Y., Chang, Y., Hu, H., Cheng, H., Huang, C., & Sun, M. (2017). Tell Me Where to Look: Investigating Ways for Assisting Focus in 360° Video. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17),* 2535-2545. New York: ACM.

MacQuarrie, A., & Steed, A. (2017). Cinematic Virtual Reality: Evaluating the Effect of Display Type on the Viewing Experience for Panoramic Video. *2017 IEEE Virtual Reality (VR)*, Los Angeles, 45-54.

Mateer, J. (2017). Directing for Cinematic Virtual Reality: How the Traditional Film Director's Craft Applies to Immersive Environments and Notions of Presence. *Journal of Media Practice 18(1)*, 14–25.

Men, L., Bryan-Kinns, N., Hassard, A. S., & Ma, Z. (2017). The impact of transitions on user experience in virtual reality. *2017 IEEE Virtual Reality (VR)*, Los Angeles, 285-286.

Moghadam, K. R., & Ragan, E. D. (2017). Towards understanding scene transition techniques in immersive 360 movies and cinematic experiences. *2017 IEEE Virtual Reality (VR)*, Los Angeles, 375-376.

Montagud, M., Boronat, F., González, J., & Pastor, J. (forthcoming). Customized and Synchronized Presentation of Subtitles in Multi-Screen Scenarios for a better QoE. Under Second Review Round, *Multimedia Systems Journal*. Springer.

Rothe, S., Heinrich, H., & Mathias, A. (2017). Diegetic cues for guiding the viewer in cinematic virtual reality. *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*. New York: ACM.

Rothe, S., Kim, T., & Hussmann, H. (2018). Dynamic Subtitles in Cinematic Virtual Reality. *Proceedings of the 15th European Interactive TV Conference (ACM TVX 2018)*. New York: ACM.

Sheikh, A., Brown, A., Watson, Z., & Evans, M. (2017). Directing attention in 360-degree video, *IBC 2016*, Amsterdam, 9-13 September.

Slater, M., & Usoh, M. (1993). Representations systems, perceptual position, and presence in immersive virtual environments. *Presence 2(3)*, 221–233.

Wang G., Gu, W., & Suh, A. (2018). The Effects of 360-Degree VR Videos on Audience Engagement: Evidence from the New York Times. In *HCI in Business, Government, and Organizations* (Nah, FH., & Xiao, B., Eds.), HCIBGO 2018. Lecture Notes in Computer Science, Vol. 10923. Cham: Springer.

Wissmath, B., Weibel, D., & Groner, R. (2009). Dubbing or Subtitling? Effects on Spatial Presence, Transportation, Flow, and Enjoyment. *Journal of Media Psychology 21(3)*, 114–125.

Witmer, B. G., & Singer, M. J. (1998). Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence Teleoperators & Virtual Environments 7(3)*, 225–240.