International Conference on Artificial Reality and Telexistence
Eurographics Symposium on Virtual Environments (2018)
G. Bruder, S. Cobb, and S. Yoshimoto (Editors)

Positioning of Subtitles in Cinematic Virtual Reality

Sylvia Rothe[1], Kim Tran[1] and Heinrich Hussmann[1]

[1]Ludwig Maximilians University Munich

Abstract

Cinematic Virtual Reality has been increasing in popularity in recent years. Watching 360° *movies with a head mounted display, the viewer can freely choose the direction of view and thus the visible section of the movie. Therefore, a new approach for the placements of subtitles is needed. In a preliminary study we compared several static methods, where the position of the subtitles is not influenced by the movie content. The preferred method was used in the main study to compare it with dynamic, world-referenced subtitling, where the subtitles are placed in the movie world. The position of the subtitles depends on the scene and is close to the speaking person. Even if the participants did not prefer one of these methods in general, for some cases in our experiments world-referenced subtitles led to a higher score of presence, less sickness and lower workload.*

CCS Concepts

• ***Human-centered computing*** → *Virtual reality;;* • ***Multimedia Information System*** → *Artificial, augmented, and virtual realities;*

1. Introduction

360° movies are attracting widespread interest and have many possible applications, e.g. telling stories about exciting locations in the world or ancient places of interest in history. Especially museums and other educational institutions can take advantage of this. In Cinematic Virtual Reality (**Cinematic VR**), the viewer watches a 360° movie using a head mounted display (**HMD**). Therefore, the viewer is inside the movie and has the possibility to look around. For watching movies in foreign languages, but also for supporting deaf viewers, subtitles are needed. Not all rules of subtitling can be transformed from traditional movies to Cinematic VR. The freedom of the viewer to choose the viewing direction requires new approaches for subtitling.

In traditional movies, usually **static** subtitles are used. These subtitles are mostly at the bottom of the movie and do not change their position. This method is also called center-bottom subtitles [KCH*17]. In Cinematic VR subtitles which are static at the bottom of the display, are not static, related to the movie world. They are connected to the display and move along with it in case the viewer is turning the head. In VR environments such objects are called **screen-referenced** [SM05, YWS99].

For reducing head and eye movements during watching traditional movies with subtitles, there are attempts to use **dynamic** subtitles placed near the speaker. The position of these subtitles is dynamically changing and depends on the scene. Other names for these subtitles are speaker following subtitles [KCH*17] or positioned subtitles [BA14]. In Cinematic VR they are connected to the virtual world, in our case to the movie, and stay fixed at their place in the movie world, when the viewer turns the head. They are **world-referenced** [SM05, YWS99].

Regarding subtitles in Cinematic VR there are three main issues. The first issue is the **position** of the subtitle. The viewer can move the head, thereby the field of view (**FoV**) is changing. There is no bottom in a 360° image, so the standard location for static subtitles is missing. Using the bottom of the display is one approach for static subtitles in Cinematic VR. World-referenced subtitles can benefit from more space between the speakers in Cinematic VR. In traditional movies there is usually only little room between dialog partners, if they are in the same shot. In other cases, only one person can be seen in one shot - the dialog partner in the next one. In Cinematic VR all talking people are on the image at the same time with some distance to each other - so the eye movements between speaking people and bottom-based subtitles are mostly greater than for subtitles placed between the speakers.

The second issue is **speaker identification**. The problem of speaker identification is more relevant in Cinematic VR than in traditional videos, as all persons in the room are visible in the 360° image at the same time, even if the viewer sees just a part of it. Placing the subtitles near the speaker, helps to identify the speaker, however the viewer is restricted in the choice of the viewing direction when reading the subtitles. In our experiments we used speaker names for the screen-referenced method and placements near the speaker for the world-referenced method to indicate the speaker.

This leads to the third issue - the **VR experience** - which includes topics such as presence, sickness and workload. Watching the movie using a HMD, the viewer gets the feeling of beeing part

DOI: 10.2312/egve.20181307

of the surrounding scenery. Since subtitles do not belong to this scenery, the presence could be reduced and additional workload or sickness could be caused.

Nowadays, in ordinary 360° movie players, e.g. the YouTube player, the subtitles are fixed at the bottom of the display (screen-referenced subtitles). Even if currently the majority of 360° videos is viewed via a flat screen, in our research we focus on subtitle methods for HMDs which needs a new approach. The findings about presence and sickness are not relevant for flat screens, however the problem of speaker identification is also present for flat screens and further research is important in this case.

Searching for a subtitling method in Cinematic VR, the following issues have to be taken into account:

- The subtitles have to be easily readable and should support the viewer's understanding of the story.
- The subtitles have to be understandable with an easy way for speaker identification.
- The subtitles should not destroy the VR experience - with as little eye strain as possible, less sickness and high presence.

Since speaker identification is an important issue for subtitling, especially in Cinematic VR, we chose scenes with more than one speaker: one dialog scene with two people and a meeting room scene with a changing number of more than three people. We compared different subtitle methods for these scenes.

As a first approach to this topic, we started studies for hearing viewers watching movies in foreign languages. We are aware of the fact that not all of our findings can be adapted to subtitles for **deaf** viewers. Even if in flat screens, subtitles for hearing and deaf people are the same, in virtual environments it makes a big difference if the viewer can hear the person - even in a foreign language - because it depends on the viewing direction if the viewer sees the speaker. A hearing person always notices, if a person starts to speak. For parts of our user study we had one deaf participant, who gave us valuable hints for our further research. We did not include this data in our analysis, as we decided to work out subtitle methods for this specific user group in the near future.

2. Related Work

2.1. Placement of Subtitles in Traditional Videos

Several studies investigated the placement of dynamic subtitles in traditional videos for reducing the switching rate and distance between regions of interests and subtitles [BA14, BJC*15, AHKM16, HWY*11, HKYW15]. Akahori et al. [AHKM16] determined the region of interest by eye tracking data of the movie and placed the subtitles at the lower part of this region to reduce eye movements. The experiments of Chen et al. [CYLJ] showed that in the case of dynamic subtitles (subtitles on the speaker's side) learners take more time to focus on the video than watching a movie with static subtitles. In our work we investigate if dynamic/world-referenced subtitles are applicable in Cinematic VR environments and if the methods have any influence on the viewing experience.

Kurzhals et al. [KCH*17] compared center-bottom subtitles with dynamic (speaker-following) subtitles in traditional videos. Dynamic subtitles led to higher fixation counts on points of interest and reduced saccade lengths. The participants had the subjective impression of understanding the content better with dynamic subtitles. In their experiments the audio was muted. Since in Cinematic VR, audio is an important cue for hearing people to recognize something new in the scene, even outside the FoV, we did not adapt this approach. Instead, we manipulated the spoken parts by reverse audio filters.

Several measuring methods for comparing cognitive load with and without subtitles in traditional movies were used by Kruger et al. [KHM13]: tracking of the pupil dilation, electroencephalography (EEG), and self-reported ratings. In their experiments same-language subtitles in an educational context reduced the cognitive load (determined by the pupil diameter) and resulted in a lower frustration level (determined by EEG).

Brown et al. [BJC*15] analyzed eye tracking data for subtitles in regular videos. They found out, that gaze patterns of people watching dynamic subtitles were more similar to the baseline than watching with traditional subtitles. Most of the participants were more immersed and missed less of the content. However, a few people preferred traditional subtitles, because they found dynamic subtitles more distracting. Another mentioned disadvantage was the fact, that for viewers who do not need subtitles, dynamic subtitles are more disruptive. This weakness is not relevant for Cinematic VR, as every viewer can choose if subtitles are desired, in contrast to traditional videos, where several people look at the same display.

2.2. Speaker Identification in Traditional Videos

Another problem besides placement of subtitles is the identification of speakers in cases where there are more than one speaker. To place the subtitles near the speaker is one of the methods which can help to solve the problem [VF10]. Other methods are: using colors, speaker names, different font types, icons and other graphical elements [VF09, VF10, K*94].

Vy and Fels [VF10] compared subtitles including speaker names with subtitles next to the speaker. In their experiments the participants felt distracted by subtitles following the speaker who changes the place. Speaker names were helpful for most participants, but not for deaf viewers, who are not aware of the voices and do not usually identify people by names, but rather by visual characteristics. A conclusion of the paper is that deaf or hard of hearing people need different methods of subtitling than hearing people. Since our participants were hearing people we used names for speaker identification in the screen-referenced method.

2.3. Subtitles in Augmented Reality

Research in Cinematic VR is very close to augmented reality (AR), where the real world takes the place of the 360°movie. Peng et al. [PHT*18] investigated speech bubbles, which are very similar to dynamic subtitles. A system with realtime speech recognition for AR-HMDs was developed for supporting deaf and hard-of-hearing people in conversations. The participants preferred speech bubbles to traditional subtitles. In our work we decided not to use speech bubbles, since they cover too much of the movie. However, the position of the subtitles and the bubbles near the speaker is the same.

Problems as speaker identification and the influence of subtitles on presence differ in AR and Cinematic VR.

2.4. Static Subtitles in 360° Videos

In their work-in-progress Brown et al. [BTP*17] suggested four static methods of subtitling for Cinematic VR (Figure 1):

- 120-Degree: the subtitles are placed at three different spots around the viewer (0°, 120°, 240°) and 15° below the horizon.
- Static-Follow: the subtitles are fixed in front of the viewer and are statically connected to the head movements. They are 15° below the 0°-line (directly ahead).
- Lag-Follow: similar to the Static-Follow method, but the subtitles remain in place for small head movements (below 30°). They are changing their position smoothly.
- Appear: each subtitle appears in front of the viewer (15° below the 0°-line). It stays static in the environment even if the viewer is moving the head. The position of the next subtitle depends on the new head direction.

We implemented these methods and compared them in a preliminary study.

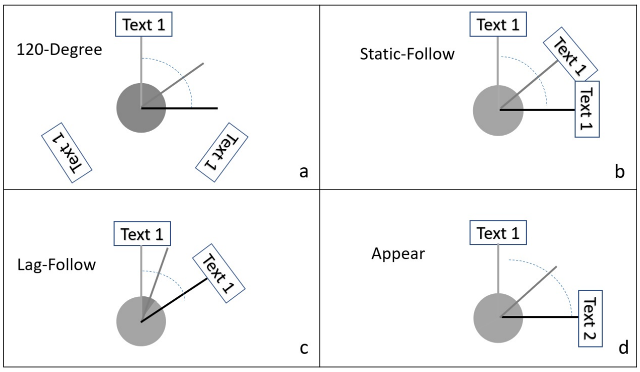


Figure 1: *Four static subtitle methods. The subtitle is*
(a) placed three times always in the same place
(b) moving with the head of the viewer
(c) moving every time, the head rotates more than 30°
(d) placed where the viewer is looking when the subtitle is changing

3. Preliminary Study

There were three goals of the preliminary study. Firstly, we were looking for the optimal font size and text placements. Secondly, we tried out a way of audio manipulation which does not confuse the viewer. The audio should sound natural, however, incomprehensible, because we wanted to motivate the user to read the subtitles. We decided not to spatialize the audio. Spatial sound could be an important aid for speaker identification, which should be investigated separately. Thirdly, we investigated which static method (Figure 1) is working best in a Cinematic VR environment. The result of the pilot study was used for the main study: the comparison with dynamic, world-referenced subtitling.

When we started our research, we wanted to compare two static methods, which were similar to the 120-Degree method and the Static-Follow method described above. Since around that time Brown et al. [BTP*17] published their proposal for four static subtitling methods, we decided to verify all four of them. The results of Brown [Bro18] were not yet published, when we finished our preliminary study. However, it turned out that the results are very similar.

3.1. Implementation

For implementing the subtitle methods, we used Unity 5.6.2 and the integrated video player. The free asset TextMeshPro was installed for modifying various text attributes of the subtitles and editing the text. Some attributes were taken from the German guidelines for subtitles in traditional movies [ARD17, Sch17]: the maximum number of characters per line (37) and the maximum permissible lines of subtitles (2). Other parameters were determined in the pilot study: the font size and distance from the horizon line. For changing the subtitles, the Animator component of Unity was used. Each of the four methods was implemented with an own script.

3.2. Material and Participants

For the pilot study, 5 voluntary participants (5 men, aged between 21 and 26) watched a short video with an Oculus Rift sitting on a swivel chair. Every participant viewed all four methods, the order of the methods was permuted. After that, in a semi-structured interview the participants were asked about the preferred method and the used parameters.

3.3. Results

All participants chose the Static-Follow method as the most comfortable and best working. Hence, for the main study we compared this method with world-referenced subtitling. For most participants 15° below the 0°-line was too far away from the viewing direction. So, we changed it to 12.5°, which was comfortable for all probands. The manipulated audio worked well.

4. Main Study

In the main study we compared screen-referenced and world-referenced subtitling. For the **screen-referenced subtitles**, the text is fixed in front of the viewer, statically connected to the HMD and so to head movements. They are 12.5° below the 0°-line. Independent of the viewing direction, the subtitles are in front - on the bottom of the display. For speaker identification, the name of the speaker is added at the beginning of the text. **World-referenced subtitles** are connected to the movie world and positioned near the speaker. It depends on the scenario where the subtitles are exactly placed.

4.1. Participants and Material

34 participants (26 men, 8 women, average age 22.9) watched the videos using an Oculus Rift. 23 participants had some experiences in VR, 19 used subtitles in their daily life. The participants saw two

short scenes recorded in a TV studio (3min length overall). In the first scene (Figure 2) two people, who do not change their places, talk to each other - we call this the "talk" video (T).

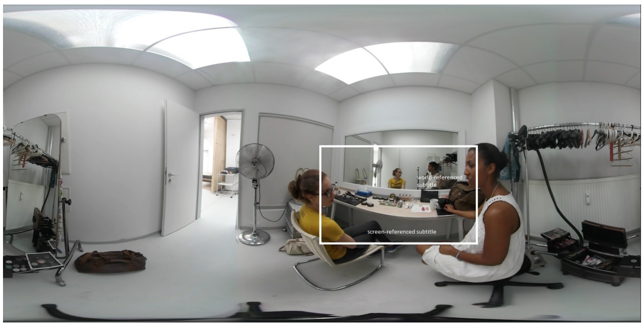


Figure 2: *The scene of the talk video (T): there are two talking persons. The frame shows the field of view, the text the positions of the subtitles, above: world-referenced, below: screen-referenced.*

In the second scene (Figure 3), there are several people in a meeting room, others are coming and leaving. This video is called the "meeting" video (M). We wanted to make sure, the participants did not understand the spoken text, therefore the audio was manipulated.

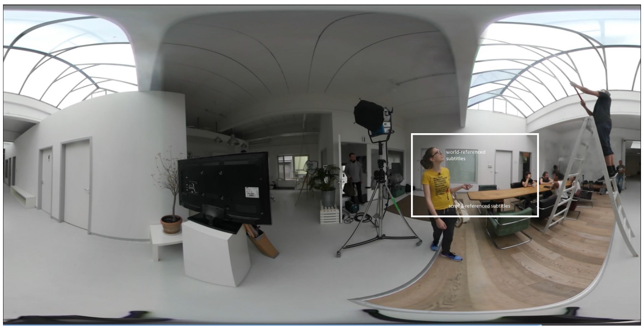


Figure 3: *The scene of the meeting video (M): several people in the same room. The frame shows the field of view, the text the positions of the subtitles, above: world-referenced, below: screen-referenced.*

4.2. Study Procedure

In case the participants had no experiences in VR, an introduction was given and a short movie was shown before the study. At the beginning, the participants were asked to fill out the first part of a questionnaire. The questionnaire consists of several segments:

- demographics (age, gender, experiences) - at the beginning
- task workload - after each video
- simulator sickness - after each video
- presence - after each video
- comparison of the two methods - at the end

After the demographics questionnaire part, every participant saw the same two short videos, each of them with one of the two methods. The order of videos and methods was counterbalanced using all four possible combinations:

- Tw - Talk with world-referenced subtitles
- Ts - Talk with screen-referenced subtitles
- Mw - Meeting with world-referenced subtitles
- Ms - Meeting with screen-referenced subtitles

So, each video/method combination was watched by 17 participants. All the head movements were tracked. After each video the task workload, sickness and presence parts of the questionnaire were answered.

Task workload: The workload was studied using the NASA-TLX questionnaire [HS88], where all six sub-scales were used: (1) Mental Demand, (2) Physical Demand, (3) Temporal Demand, (4) Performance, (5) Effort, (6) Frustration.
As described in [Har06] we eliminated the weighting process and used the Raw TLX (RTLX). In addition to the overall load, the subscale rates of each single item were compared for finding possible reasons for increased workload.

Simulator sickness: For measuring simulator sickness a reduced questionnaire of the Simulator Sickness Questionnaire (SSQ) of Kennedy et al. [KLBL93] was used. Since not all questions are relevant for Cinematic VR, six items were selected: (1) general discomfort, (2) fatigue, (3) headache, (4) eye strain, (5) difficulty focusing, (6) nausea, (7) difficulty concentrating.
In this way it was not possible to calculate the total score exactly as it is described in the original paper [KLBL93]. However, using the selected items, we could calculate scores for nausea, oculomotor and disorientation to compare both methods. Additionally, we inspected the rates of each item in detail to find possible reasons for sickness. Similar to the task workload evaluation we used raw scores without weighting [BSJRW09].
For each item one of the sickness levels (none, slight, moderate, severe) could be chosen and the answers were transformed to a scale from 0 (none) to 3 (severe). The reported rates are the sums of these values [KLBL93].

Presence: To investigate the presence, we used parts of the presence questionnaire (PQ) of Witmer and Singer [WS98]. Since the PQ was developed for general virtual environments with interactivity and movement, we chose some of the questions which are relevant for Cinematic VR:

- How involved were you in the virtual environment experience? (involvement)
- How much did the visual aspects of the environment involve you? (visual)
- How much did the auditory aspects of the environment involve you? (audio)
- How quickly did you adjust to the virtual environment experience? (time)
- How much did your experiences in the virtual environment seem consistent with your real-world experiences? (accordance)

After each video a semi-structured interview was held and recorded. This interview was based on the questions in [BJC*15] with some additional issues about the content. The questionnaire ended with some questions comparing the two methods.

5. Results

5.1. Analysis of the Spatiotemporal Data

In our experiments we collected two types of data: the head orientation tracking coordinates (spatiotemporal data), and the answers of the questionnaires. Inspecting the head tracking data, we could not find significant differences between the two methods for the **meeting video**. The hotspots were distributed in the room - not only around the speaking persons.

However, in the dialog of the **talk video** the participants were more focused on the speakers when the world-referenced method was used (Figure 4).

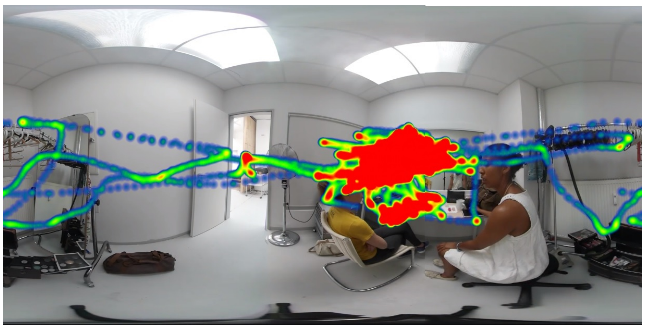


Figure 4: *Heatmap of the head tracking data for the talk video with world-referenced subtitles (Tw). There is a cluster in the area of the subtitles.*

The participants who watched this scene with the screen-referenced method looked around more often during the dialog of the protagonists (Figure 5).

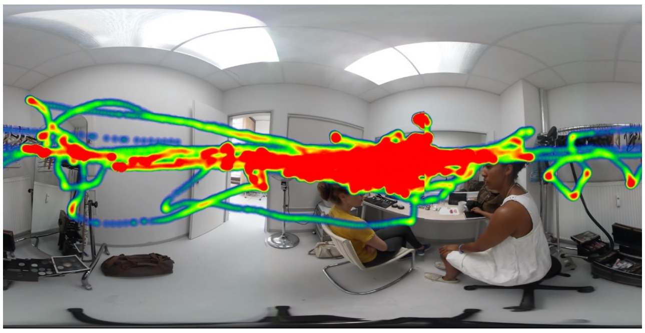


Figure 5: *Heatmap of the head tracking data for the talk video with screen-referenced subtitles (Ts). The viewers are looking more around than in the world-referenced case.*

With the world-referenced method the user has to stay near the speakers for reading the subtitles. In contrast, the screen-referenced method, where the subtitles are fixed on the display, enables the viewer to look around and read at the same time.

5.2. Analysis of the Questionnaires

In the answers of the comparison part of the questionnaire we did not find preferences for one of the methods. However, comparing the score of the sickness and workload parts, the viewers felt in some aspects more comfortable with the world-referenced method. Additionally, the presence questions were answered with a slightly higher rate for the world-referenced method.

Task Workload: Comparing the total scores for NASA-TLX, the workload for nearly all items, except physical demand and performance, was higher for the screen-referenced method. (Table 1).

	Screen-referenced	World-referenced	p-value (t-test)
Mental Demand	**59.41** (27.44)	**43.09** (26.54)	.007
Physical Demand	24.27 (17.84)	31.32 (24.26)	.08
Temporal Demand	**57.35** (30.23)	**45.15** (26.04)	.03
Performance	45.74 (26.06)	44.71 (26.97)	.4
Effort	**48.82** (28.04)	**36.32** (22.47)	.02
Frustration	**46.32** (29.42)	**30.15** (22.21)	.006
Total Demand	46.98 (20.23)	38.45 (15.48)	.38

Table 1: *Rates for several sections of the NASA-TLX (means, standard deviation). The significant differences between the two methods were evaluated by t-tests and marked bold. In all of these cases the rates are higher for the screen-referenced method.*

In a closer inspection we investigated the data for the two videos separately (Figure 6). We performed a two sample t-test per item, which did not show a significant difference for nearly all items in the **meeting** video. However, there is a higher workload score for the world-referenced method in physical demand (t-test: t=-3.35, df=16, p=.001).

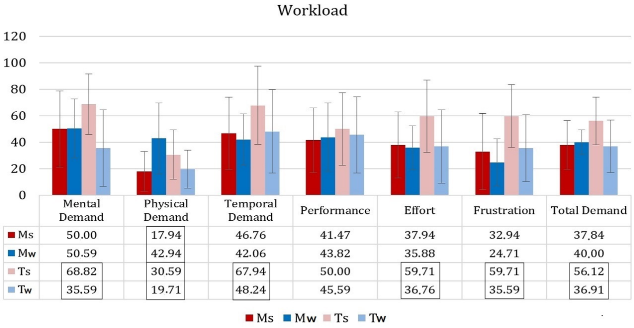


Figure 6: *Scores of the NASA-TLX questionnaire (means, standard deviation) for every video/method combination. The significant differences areas are marked by a border.*

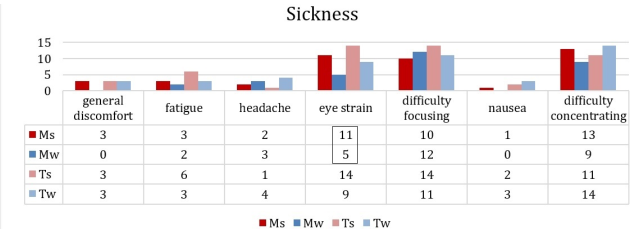
Inspecting the data of the **talk** video and perfoming t-tests for each item, we found significant higher workload for the screen-referenced method in:
- mental demand (t = 3.7, df = 16, p = .0004).
- physical demand (t = 1.9, df = 16, p = .032)
- temporal demand (t = 1.8, df = 16, p = .035)
- effort (t = 2.4, df = 16, p = .01)
- frustration (t = 2.9, df = 16, p = .004)
- in total demand (t = 2.9, df = 16, p = .003)
The differences for all other items are not significant.

Simulator Sickness: Generally, the participants only had slightly noticeable discomfort. However, the rates for all sections of the SSQ were higher for the screen-referenced method (Table 2). Since we did not use all the questions of the questionnaire, the scores in Figure 7 cannot be compared to results of other papers. They were only used to compare both methods.

	nausea	oculomotor	disorientation
screen-referenced	33	91	24
world-referenced	29	75	23

Table 2: *Rates for several sections of the SSQ. In all cases the rates are higher for the screen-referenced method.*

Considering the questions separately, for most items the difference was slight, however the eye strain score was significant higher for the screen-referenced method (Figure 7) (t-test: t=1.7, df=16, p=.047).



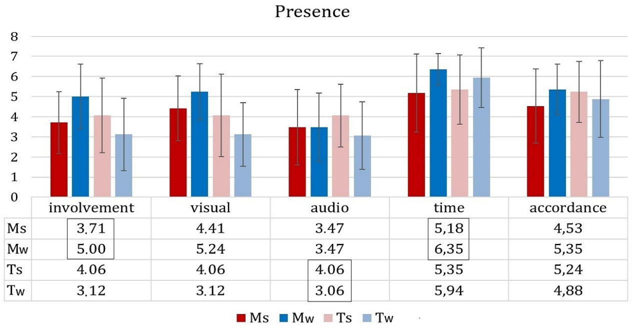
	general discomfort	fatigue	headache	eye strain	difficulty focusing	nausea	difficulty concentrating
Ms	3	3	2	11	10	1	13
Mw	0	2	3	5	12	0	9
Ts	3	6	1	14	14	2	11
Tw	3	3	4	9	11	3	14

Figure 7: *Rates (summations) of Simulator Sickness for every video/method combination. There is only a significant difference for eye strain.*

Presence: Comparing the answers to the PQ for both methods, the differences were slight in nearly all cases. However, the question "How quickly did you adjust to the virtual environment experience?" resulted in a significant higher score for the world-referenced method (t-test: t = -2.3, df = 33, p = .01) (Table 3).
To find more detailed results, the answers for the two different videos were analyzed separately (Figure 8) and a two sample t-test was performed for every item. We found significant higher presence levels in the following combinations:
- Ts: audio item (t=1.8, df=16, p = .041)
- Mw: involvement item (t=-2.39, df=16, p= .011)
- Mw: time (t=-2.31, df=16, p = .014).

	Screen-referenced	World-referenced	p(t-test)
How involved ...?	3.88 (1.68)	4.06 (1.94)	0.3
... visual aspects ...?	4.24 (1.83)	4.18 (1.82)	0.4
... auditory aspects ...?	3.67 (1.72)	3.26 (1.68)	0.1
How quickly ...?	**5.26(1.81)**	**6.15 (1.18)**	0.01
How ... consistent ...?	4.88 (1.70)	5.12 (1.61)	0.2

Table 3: *Means and standard deviation for the presence questions. Significant differences between the two methods are marked bold.*



	involvement	visual	audio	time	accordance
Ms	3.71	4.41	3.47	5,18	4,53
Mw	5.00	5.24	3.47	6,35	5,35
Ts	4.06	4.06	4.06	5,35	5,24
Tw	3.12	3.12	3.06	5,94	4,88

Figure 8: *Means and standard deviations for the presence questions for every video/method combination. Bordered values show significant differences between the two methods.*

In Table 4 the results of the questionnaire analysis are summarized.

	Talk	Meeting	Overall
Task Workload (lower)	world	physical: screen	world
Sickness (lower)		eye strain: world general discomfort:world	world
Presence (higher)	audio:screen	involvement:world time:world	time:world

Table 4: *Summary of the results. The table shows the methods with the best result.*

5.3. Qualitative Analysis

After every section of the questionnaire the participants could give some comments about the methods. Additionally, a semi-structured interview was carried out. In summary, the following statements were mentioned.

Screen-referenced Method, positive:

- "I can decide where to look." (P10, P30, P11, P19, P23, P31, P34, P16, P32)
- "This method is similar to the method in TV." (P18)
- "The subtitles are always visible." (P5, P6, P2, P26)

Screen-referenced Method, negative:

- "It is difficult to assign the speaking person." (P14)

World-referenced Method, positive:

- "Subtitles can be assigned more easily to the speaker." (P1, P25, P14, P3, P15, P8, P24)
- "Speakers and subtitles can be seen simultaneously." (P33)
- "It is a more natural experience. " (P13, P15, P12, P24)
- "It is easier to absorb the content." (P17, P21, P4, P8)

World-referenced Method, negative:

- "I am forced to look at the speaker." (P7, P19)
- "It is sometimes difficult to discover the speaker." (P29, P33)
- "I did not know where the next subtitle will appear." (P33)

Asking the participants directly about the preferences for the two subtitling methods, the results were well-balanced. In contrast to the NASA-TLX, SSQ and PQ parts we could not find a preferred method. However, we got important hints for problems which should be solved in the future.

6. Discussion and Limitation

In this work we explored several methods of subtitling. First the **position** of static subtitles was tested. Comparing four techniques, the Static-Follow method, where the subtitles are connected to the display and thus to head movements, was preferred by all users of our pilot study. This result coincides with the study of the BBC research department [Bro18]. In a second step this method was compared with subtitles which are connected to the movie world. We could not find one generally preferred method. The position of screen-referenced subtitles makes it easier to look around, but the participants had the subjective impression that it is more difficult to absorb the content.

For **speaker identification** the world-referenced method is preferred. The viewer can see the speaking person and read the subtitles simultaneously without extensive eye movements. However, if the speaking person is changing and the following person is not in the FoV, it needs some effort to find the new speaker and subtitle. If there are more than one speaking person in the movie, it is difficult to match the subtitles to the speakers using the screen-referenced method. So, it is more difficult to understand the story.

At the moment the most CVR movies do not use spatial sound. Also, in our user study the sound was not spatial. Spatial sound can be an important aid in speaker identification for hearing people. This should be further investigated.

Even if the participants did not prefer one of the methods in the comparison part of the questionnaire, the questions about the **VR experience** result in better scores for the world-referenced method. One reason could be that world-referenced subtitles are integrated in the movie and screen-referenced subtitles are part of the display. Participants noted, that world-referenced subtitles are "more natural, it coincides more with the real life". Comparing the data regarding task workload, sickness and presence the world-referenced subtitling method was more comfortable in several cases. There is less eye strain because the subtitles are placed near the speakers and the viewer is not forced to switch to the bottom of the FoV. The screen-referenced method has a better score just in one item: physical demand in the meeting scene.

In our experiment the screen-referenced method led to higher **workload** in most cases. One explanation for this could be that following subtitles and watching the movie need more effort if speaker and text are not linked. One exception is the higher physical demand for the world-referenced method in the meeting scene. This could be caused by the fact that the people are not close to each other, so the viewer has to search the next person for reading the subtitle and also to move the head.

There was only slight **sickness** during the experiments. For most items the difference between the methods is small, however the eye strain score is higher for the screen-referenced method (Figure 7), caused by more eye movements for switching between the person and the text.

Analyzing the **presence** questions for the meeting video, the world-referenced method resulted in higher scores for the questions "How involved were you in the virtual environment experience?" and "How quickly did you adjust to the virtual environment experience?" (Figure 8). A reason for that could be that world-referenced subtitles tend to be more part of the environment than subtitles connected to the display.

The results of our study depend on the type of the scene. We explored two types of scenes: a dialog of two persons and a group of speaking people. The protagonists did not change their positions during the conversation. For moving protagonists, who are speaking, world-referenced subtitles need to move accordingly, which could cause sickness. Such scenarios require further testing.

Inspecting the heatmaps of the head tracking data, we found differences for the talk scene. In time intervals where people were speaking, the data of the world-referenced methods are more concentrated around the speakers, which means less head movements. This could be one reason for the lower task workload which results from the answers of the NASA-TLX.

The participants of this study were hearing people. So, the results can be helpful for finding subtitle methods for foreign languages. It requires more experiments to find out if deaf or hard of hearing viewers need other approaches.

For logging the viewing direction, we used head tracking. The additional usage of an eye tracker could lead to more detailed results in the analysis of the viewing direction.

The video material of our user study was very short (overall 3min). However, the participants reported some discomfort. Reading subtitles in Cinematic VR for a longer time needs further research.

32% of our participants were beginners in VR and reading subtitles in cinematic VR was new for everybody. Doing this more often and being more familiar with the use of subtitles can lead to less effort and more comfortable reading.

7. Conclusion and Future Work

Both methods - screen-referenced and world-referenced subtitling - are helpful for understanding movies in foreign languages. Even if our work is just a first approach and we investigated just two special scenes, the results of this study encourage further studies in this field. We think there is much potential in world-referenced subtitles which are not used in Cinematic VR at the moment. However, none of the investigated methods meet all requirements for each scenario in Cinematic Virtual Reality. A combination of the methods depending on the requirements could be a new approach: the world-referenced subtitles are used when the speaker is in the field of view. They switch to screen-referenced subtitles when the viewer turns the head and the speaker disappears from the FoV. Additionally, the subtitling methods could be expanded with techniques of attention guiding to facilitate speaker identification.

Furthermore, we will continue our work with deaf or hard of hearing people, where more effort in the speaker identification is needed. For hearing people, the voices of the protagonists are an aid which is not available for deaf people. Different colors, fonts or signs are already used in subtitling of traditional movies and could be adapted. However, the problem of speaker identification in cinematic VR is harder than in traditional movies and needs further research.

Even if we are just at the beginning of finding useful subtitle methods for Cinematic VR, these techniques are also important in other areas such as Augmented Reality and other fields of virtual reality.

References

[AHKM16] Akahori W., Hirai T., Kawamura S., Morishima S.: Region-of-interest-based subtitle placement using eye-tracking data of multiple viewers. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video* (2016), ACM, pp. 123–128. 2

[ARD17] ARD: *Untertitelstandards*, (accessed November 13, 2017). http://www.daserste.de/service/kontakt-undservice/ barrierefreiheit-im-ersten/untertitel-standards/index.html. 3

[BA14] Brooks M., Armstrong M.: Enhancing subtitles. *TVX2014 Conference, Brussels* (2014), 25–27. 1, 2

[BJC*15] Brown A., Jones R., Crabb M., Sandford J., Brooks M., Armstrong M., Jay C.: Dynamic subtitles: the user experience. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video* (2015), ACM, pp. 103–112. 2, 4

[Bro18] Brown A.: *Exploring Subtitle Behaviour for 360° Video - BBC R&D*, 2018 (accessed July 13, 2018). https://www.bbc.co.uk/rd/publications/whitepaper330. 3, 7

[BSJRW09] Bouchard[1] S., St-Jacques[1] J., Renaud[1] P., Wiederhold B. K.: Side effects of immersions in virtual reality for people suffering from anxiety disorders. *Journal of CyberTherapy & Rehabilitation 2*, 2 (2009). 4

[BTP*17] Brown A., Turner J., Patterson J., Schmitz A., Armstrong M., Glancy M.: Subtitles in 360-degree video. In *Adjunct Publication of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video* (2017), ACM, pp. 3–8. 3

[CYLJ] Chen H., Yan M., Liu S., Jiang B.:. 2

[Har06] Hart S. G.: Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (2006), vol. 50, Sage Publications Sage CA: Los Angeles, CA, pp. 904–908. 4

[HKYW15] Hu Y., Kautz J., Yu Y., Wang W.: Speaker-following video subtitles. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 11*, 2 (2015), 32. 2

[HS88] Hart S. G., Staveland L. E.: Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology 52* (1988), 139–183. 4

[HWY*11] Hong R., Wang M., Yuan X.-T., Xu M., Jiang J., Yan S., Chua T.-S.: Video accessibility enhancement for hearing-impaired users. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 7*, 1 (2011), 24. 2

[K*94] King C. M., et al.: Digital captioning: Effects of color-coding and placement in synchronized text-audio presentations. 2

[KCH*17] Kurzhals K., Cetinkaya E., Hu Y., Wang W., Weiskopf D.: Close to the action: Eye-tracking evaluation of speaker-following subtitles. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), ACM, pp. 6559–6568. 1, 2

[KHM13] Kruger J.-L., Hefer E., Matthew G.: Measuring the impact of subtitles on cognitive load: Eye tracking and dynamic audiovisual texts. In *Proceedings of the 2013 Conference on Eye Tracking South Africa* (2013), ACM, pp. 62–66. 2

[KLBL93] Kennedy R. S., Lane N. E., Berbaum K. S., Lilienthal M. G.: Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology 3*, 3 (1993), 203–220. 4

[PHT*18] Peng Y.-H., Hsi M.-W., Taele P., Lin T.-Y., Lai P.-E., Hsu L., Chen T.-c., Wu T.-Y., Chen Y.-A., Tang H.-H., et al.: Speechbubbles: Enhancing captioning experiences for deaf and hard-of-hearing people in group conversations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), ACM, p. 293. 2

[Sch17] Schneider. B.: *Gemeinsame Untertitelrichtlinien für den deutschen Sprachraum.*, (accessed November 13, 2017). http://www.untertitelrichtlinien.de. 3

[SM05] Shah P., Miyake A.: *The Cambridge handbook of visuospatial thinking*. Cambridge University Press, 2005. 1

[VF09] Vy Q. V., Fels D. I.: Using avatars for improving speaker identification in captioning. In *IFIP Conference on Human-Computer Interaction* (2009), Springer, pp. 916–919. 2

[VF10] Vy Q. V., Fels D. I.: Using placement and name for speaker identification in captioning. In *International Conference on Computers for Handicapped Persons* (2010), Springer, pp. 247–254. 2

[WS98] Witmer B. G., Singer M. J.: Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and virtual environments 7*, 3 (1998), 225–240. 4

[YWS99] Yeh M., Wickens C. D., Seagull F. J.: Target Cuing in Visual Search: The Effects of Conformality and Display Location on the Allocation of Visual Attention. *Human Factors: The Journal of the Human Factors and Ergonomics Society 41*, 4 (dec 1999), 524–542. doi:10.1518/001872099779656752. 1