Close to the Action: Eye-Tracking Evaluation of Speaker-Following Subtitles

Kuno Kurzhals[1], Emine Cetinkaya[1], Yongtao Hu[2], Wenping Wang[2], Daniel Weiskopf[1]

[1]Visualization Research Center (VISUS)
University of Stuttgart, Germany
{kuno.kurzhals, weiskopf}@visus.uni-stuttgart.de

[2]Department of Computer Science
The University of Hong Kong, Hong Kong
{ythu, wenping}@cs.hku.hk

ABSTRACT

The incorporation of subtitles in multimedia content plays an important role in communicating spoken content. For example, subtitles in the respective language are often preferred to expensive audio translation of foreign movies. The traditional representation of subtitles displays text centered at the bottom of the screen. This layout can lead to large distances between text and relevant image content, causing eye strain and even that we miss visual content. As a recent alternative, the technique of speaker-following subtitles places subtitle text in speech bubbles close to the current speaker. We conducted a controlled eye-tracking laboratory study ($n = 40$) to compare the regular approach (center-bottom subtitles) with content-sensitive, speaker-following subtitles. We compared different dialog-heavy video clips with the two layouts. Our results show that speaker-following subtitles lead to higher fixation counts on relevant image regions and reduce saccade length, which is an important factor for eye strain.

ACM Classification Keywords

H.5.1. Information Interfaces and Presentation (e.g., HCI): Multimedia Information Systems; H.5.2. User Interfaces: Screen Design (e.g., text, graphics, color)

Author Keywords

Eye tracking; video; subtitle layout

INTRODUCTION

Subtitles play an important role in communicating content in media such as movies and TV shows [19]. They are the most prominent solution to substitute audio for hearing-impaired persons. On public displays, especially in noisy environments (e.g., subways), subtitles are applied to convey information without disturbing people with audio. Subtitles are also an efficient approach to translating and communicating audio-visual content. Professional audio translation and synchronization (dubbing) is expensive and time-consuming. In many countries, the original audio is kept and only subtitles in the local language are added. Even when applied in the same language as the audio, subtitles can help improve reading skills [20].

CHI 2017, May 06–11, 2017, Denver, CO, USA

ACM 978-1-4503-4655-9/17/05...$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025772

Nowadays almost everyone who is interested in creating subtitles can do so. As an example, *YouTube*[1] provides the possibility to contribute subtitles in any language for the videos on the platform. This provides the possibility for users to reach a worldwide audience with the help of the viewer community.

In summary, creating subtitles is an efficient and simple way to translate and present multi-language information in multimedia content. However, the regular approach shows subtitles at the center-bottom of the screen. This layout has the disadvantage that the visual angle between subtitles and the actual image content is increased artificially. With the tendency to read text even when the audio language is known [11], the viewer has to foveate the respective region at the bottom of the screen to read. With an approximated viewing angle of 2° for foveal vision [31], and a rapid decrease of visual acuity between 2°–6°, the viewer has to keep switching the gaze position between text and image. Although people are cognitively able to process text and image content together [28], this can lead to increased eye strain. Also, content, either in the image or in the subtitles, might be missed [5].

This issue can be addressed by subtitles with dynamically changing positions, sensitive to the presented content. The *speaker-following subtitle* technique [17] transforms regular subtitle text into speech bubbles (similar to those in comics), close to the speaker's face. This approach can be interpreted as an application of Fitts' law [12, 26] to the subtitles and the image. The algorithm is also sensitive to occlusions of other persons and optimizes the position of the text. Decreasing the distance between text and image elements should make it easier to switch between them and improve the viewing experience.

To this point, the evaluation of subtitles was often restricted to traditional subtitles with a center-bottom layout. Alternative layouts were presented, but their evaluation was mainly restricted to subjective measures where participants filled out questionnaires about their impressions and experiences. With eye tracking, it is possible to record a viewer's eye movements and apply different metrics [15, 29] in order to quantify viewing behavior objectively.

Our main contribution is a comparative evaluation of participants' viewing behavior for watching video with subtitles. We compare between regular subtitles and the speaker-following subtitle layout. With eye tracking as objective and a questionnaire as subjective measure, we can, on the one hand, quantify

[1]www.youtube.com

different viewing behavior and, on the other hand, evaluate how well the layouts are accepted. Furthermore, we provide a visual analysis of the recorded data that provides further details about the spatio-temporal gaze distribution for the investigated subtitle layouts.

RELATED WORK

For traditional subtitles, many different aspects to represent and formulate subtitle texts have been discussed and evaluated (e.g., Chiaro et al. [8], Karamitroglou [18], and Koolstra et al. [19]). Our focus is on related work that included eye-tracking measurements to evaluate human viewing behavior on subtitles.

Szarkowa et al. [33] compared verbatim, standard, and edited subtitle texts with hearing and hearing-impaired participants. For representation of the text, the regular layout (center-bottom) was used. The authors found that dwell time on subtitles can be reduced when edited or standard text is used. Krejtz et al. [21] evaluated the influence of shot changes on the reading behavior of subtitles. Their results did not support the assumption that shot changes induce re-reading of text. They also included hearing and hearing-impaired participants in their study. Rajendran et al. [30] investigated the influence of different text chunking strategies for subtitles. The authors found that text chunking by phrase or by sentence reduces the amount of time spent on subtitles, and text can be processed more easily.

Among other measures, Kruger et al. [22, 23] applied eye tracking (pupil dilation) to measure cognitive load when watching an academic lecture with and without subtitles. Their results supported the use of subtitles in an educational setting to reduce cognitive load. However, their work did not focus on the distribution of visual attention. d'Ydewalle et al. [10, 11], Bisson et al. [6], and Ross and Cowler [32] conducted experiments that showed that participants tend to read subtitles, even if the language of the audio is known. Their results also indicate that this tendency exists for audio and non-audio conditions. We excluded the audio from our experiment to focus on the visual components.

Perego et al. [28] conducted a user study to analyze the effectiveness of subtitle processing. The authors observed no tradeoff between text recognition and scene recognition from a cognitive point of view. Nevertheless, the incorporation of subtitles leads to changes in the viewing behavior and the comparative studies mentioned next, show that from a viewer's perspective, the individual viewing experience is also influenced by subtitles.

All these studies focused on aspects of regular subtitles. A comparison between different position layouts was not performed. Although different layouts have been proposed over the years, evaluation focused more on the precision of the algorithms and subjective impressions of the users.

Hong et al. [16] presented and evaluated an algorithm that calculated new subtitle positions to improve the viewing experience of hearing-impaired people. They created content-sensitive subtitles that also displayed variations in the voice volume. They asked the participants to rate *enjoyment* and *naturalness* of dynamic and regular static subtitles. A majority of the hearing-impaired participants preferred their approach. The participants who preferred the static approach justified their choice with their acquaintance to regular subtitles.

Brown et al. [7] conducted a user study comparing regular and dynamic subtitles. The authors focused on the qualitative evaluation of user experience when watching videos with the different subtitle layouts, based on self-reported data. They also included an evaluation of recorded eye-tracking data, measuring the deviation of gaze behavior in comparison to the baseline of videos without subtitles. The main result of their eye-tracking study was the fact that dynamic subtitles create gaze patterns closer to the baseline than regular subtitles. We partially reproduce and confirm their results. On top of that, our study extends theirs by investigating how the gaze patterns for different subtitle layouts look. We focus our experiment on visual aspects alone, removing audio as a potential influence factor. We apply established eye-tracking metrics (i.e., fixation count, fixation duration, and saccade length) on the gaze data for inferential statistics (hypothesis testing). Whereas Brown et al. computed the difference of gaze distributions on regular grids, we apply hypothesis testing based on predefined areas of interest (AOIs), which incorporate semantic information about the stimulus. We will detail the difference between both measurement approaches in our discussion section. Furthermore, we contribute a visual analysis of the recorded eye-tracking data: We inspect the overall gaze distribution in a spatio-temporal overview using a space-time cube.

Our user study compares regular subtitles with speaker-following subtitles, according to Hu et al. [17]. In their work, the authors evaluated their approach with an online user study ($n = 219$). Participants rated regular subtitles, speaker-following subtitles, and the alternative layout by Hong et al. [16] considering the overall viewing experience and the subjective degree of eye strain. In summary, Hu et al.'s approach was rated better than the alternatives in overall viewing experience and it reduced eye strain. The authors did not include any evaluation of the viewing behavior. Due to the good results of their approach, we included speaker-following subtitles in our comparative eye-tracking study to further evaluate how the layout influences the viewing behavior.

SPEAKER-FOLLOWING SUBTITLES

In contrast to regular subtitles that typically layout text in a static area, the concept of speaker-following subtitles focuses on the integration of text closer to the relevant image region (i.e., the face of the speaker). The algorithmic concept of this approach is described in detail by Hu et al. [17]. To facilitate the understanding of how the algorithm works, we summarize the main steps to create content-sensitive speech bubbles from regular subtitle text (Figure 1):

- **Segmentation:** First, the video is segmented according to the timestamps in the subtitle file. With this step, *speaking* and *non-speaking* segments can be distinguished.
- **Speaker Detection:** To identify speaking persons, face detection and face tracking are preformed. The resulting

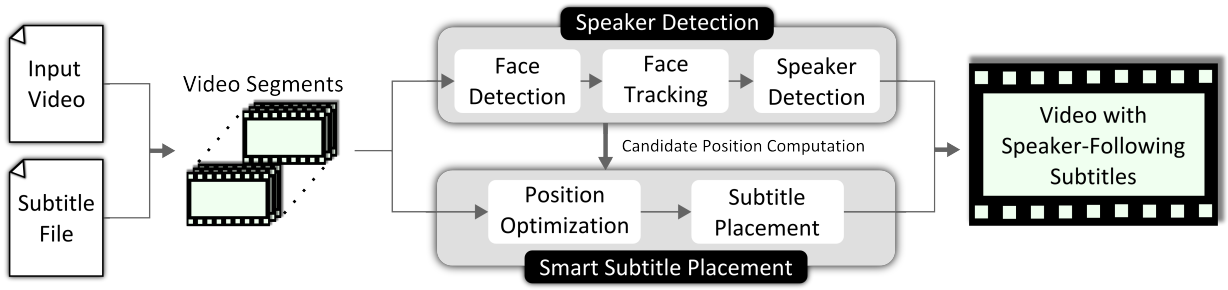


Figure 1: Processing steps to create speaker-following subtitles, according to Hu et al. [17]. The two main components are *speaker detection* and *smart subtitle placement*.

(a) Regular subtitle

(b) Speaker-following subtitle

Figure 2: Comparison between the two layouts for subtitles: (a) Regular subtitles are typically placed at the center-bottom of the screen. (b) Speaker-following subtitles are content-sensitive and appear close to the face of the corresponding speaker. Please note that video images are shown in a stylized form in the figures.

tracklets are then analyzed with respect to different features (e.g., lip motion) to identify the current speaker. With this approach, the identification of speaking persons is restricted to on-screen presence of the speaker. For off-screen dialog (e.g., from a narrator), no changes to subtitles are performed.

- **Smart Subtitle Placement:** According to the issues with regular subtitles, the new layout has to place subtitles close to the speaker, away from image borders, without occluding other important visual content, and being time-consistent with previous subtitle positions. Eight *candidate positions* are calculated around the speaker's face. From these positions, an optimal position is selected, according to the mentioned constraints.

The maximum width of a text line is restricted to a third of the image width, as determined by Hu et al. [17] in previous experiments. Line breaks are not allowed to happen within a word, preferably at punctuation marks and spaces. In general, this leads to an improved aspect ratio for long text segments, which shows text more compactly than by stretching it along a horizontal line.

The final representation of the subtitles shows black text with a white halo in a transparent rectangle (Figure 2). A small arrowhead on the rectangle indicates the position of the speaker. For both layouts, a fixed font size of 20 pt was applied for rendering the text.

USER STUDY

For the comparison between regular and speaker-following subtitles, we conducted a user study in a mixed design (within/between subjects) with different example videos consisting of dialog scenes. As objective measurement, we applied eye tracking to capture fixations and saccades of the participants. To evaluate subjective impressions of the participants, we included a questionnaire.

Hypotheses

Based on the measured eye-tracking data from a pilot study, we hypothesize that the subtitle layout has a significant effect on the viewing behavior of the participants. With the optimized position of the subtitles in the video, participants can switch between text and image content more efficiently. This results in a better distribution of attention and shorter saccades for a less exhausting viewing experience. In detail, we formulate six hypotheses considering the distribution of attention and saccade changes:

- **H1: The average fixation count with regular subtitles is higher than with speaker-following subtitles.**

  In our pilot study, we noticed more fixations with the regular subtitles. This could be a result of the different text chunking in the layouts, similar to the effects described by Rajendran et al. [30].

- **H2: The average fixation duration with speaker-following subtitles is higher than with regular subtitles.**

  Due to the improved aspect ratio of the speech bubbles and the optimized position, the participants can foveate more text than on a horizontal line in the regular layout. Additionally, a higher mental workload can lead to shorter fixation durations [15]. Along with **H1**, a higher fixation count and shorter fixation durations could support the assumption that the mental workload is higher on regular subtitles.

- **H3: The average saccade length for regular subtitles is higher.**

  This results from the distance between text and image. When the participants switch between text and image, they have to overcome longer viewing distances. Longer saccades (increased amplitude) are an important factor to cause fatigue effects [2].

- **H4: The average fixation count on faces is higher with speaker-following subtitles.**

  With subtitles being close to the speaker, the participants can better focus on the image content.

- **H5: The average fixation count on text is higher with regular subtitles.**

  Participants have to read text on a horizontal line. Compared to the compact representation of text in speech bubbles, they have to switch the foveated area more often.

- **H6: The average transition count between text and faces is higher for speaker-following subtitles.**

  With the optimized layout and the shorter distance between text and image content, the participants can switch more often between these AOIs. According to Holmqvist et al. [15], "when an area is so important that it must be more or less continuously monitored, transition rates drop". We assume that by optimizing the layout, we can reduce the importance of the subtitles, allowing the participants to switch more often to the image.

Stimuli and Task

We selected a set of ten videos for presentation (Table 1). All stimuli show dialog scenes with two or multiple persons talking to each other. We selected these videos because they were used by Hu et al. [17] to evaluate the algorithm to create the speaker-following subtitles. The layouts were created with German subtitles because all participants spoke German as first or second language. We presented each participant all videos once, switching between regular and speaker-following subtitles after each video. The order of the videos and of the subtitle layout was counter-balanced using Graeco-Latin squares. All videos were resized to a uniform height of 720 pixels, the width was adjusted according to the aspect ratio of the individual videos.

The participant's task was to attend the video and summarize its content after the video was over. This task served as motivation to read the subtitles and all participants were able to recapitulate the content in 2–3 sentences. To ensure that all participants would solve the task by reading the subtitles and not by listening, we removed the audio track from the videos. With this step, we can rule out that participants ignored the subtitles because they understood the language of the audio. We also asked the participants after each video if they had seen it before. Although some participants were familiar with some of the clips, we could not find any significant differences considering the fixation of subtitles. Hence, the influence of the participants' knowledge of the movie can be neglected in this case.

Table 1: List of video stimuli consisting of short clips from movies and TV shows.

ID	Video	Duration (min:sec)
1	Erin Brokovich 1	2:03
2	Erin Brokovich 2	2:12
3	Friends	1:21
4	Kramer vs Kramer	1:40
5	Lions for Lambs	2:34
6	Scent of a Woman	2:36
7	The Big Bang Theory	1:38
8	The Man from Earth	2:34
9	Up in the Air 1	2:48
10	Up in the Air 2	2:44

Technical Setup

For the experiment, we used a Tobii T60XL remote eye tracker with a 24" screen (resolution 1920×1200). We used a chin rest to fixate participants at an eye distance of 65 cm from the screen. A nine-point calibration was initially performed. The recorded data was preprocessed with the Tobii fixation filter (velocity threshold = 35 pixels/samples; distance threshold = 35 pixels). We approximate saccade lengths as the distances between consecutive fixations from the filtering step. Because the presented videos contain mainly dialog scenes, smooth pursuit eye movement was neglected.

Pilot Study

We conducted a pilot study ($n = 11$) to test the study design and identify flaws. As a result, we identified an effect of the algorithm that had to be compensated for the final study: If no speaker can be identified, the subtitle will be placed at the center-bottom. When the speaker can be identified during the timespan the text is visible, the algorithm moves the text to the optimized position. This can lead to "jumping" subtitles, which was noted as stressful by the participants. To address this issue, we identified occurring jumps of subtitles and replaced them by regular subtitles. For the statistical evaluation, the corresponding frames (approximately 15% per video) were removed from the data.

Participants

The user study was conducted with 40 participants (17 female, 23 male) with an average age of 23 years. The youngest participant was 18 years old, the oldest participant 33 years. All participants had an academic background. 14 participants had corrected to normal eye sight. A Snellen and an Ishihara test were conducted to test the participants' eye sight and

Table 2: Overview of the eye-tracking results. A (*) in the last column indicates a significant difference ($p < 0.01$).

Hypothesis	Measure	Test
H1	Fixation count	(U-test)
H2	Fixation duration	(t-test)
H3	Saccade length	* (t-test)
H4	Fixations on faces	* (U-test)
H5	Fixations on subtitles	* (U-test)
H6	Transition count	(t-test)

color vision. All participants passed the tests. The data from three participants had to be replaced due to insufficient eye-tracking data. A participant's eye-tracking data was considered insufficient when more than 25% of the sampled data was discarded by the eye tracker.

Study Procedure

The participants were asked to sign a consent form and to fill out a questionnaire that included information about their gender, age, academic background, and how often they watch video content with subtitles. Then, we performed the aforementioned vision tests. In a short tutorial, the two different layouts were introduced and the differences explained. For the tutorial, an additional video was prepared that was not included in the final set of stimuli. Then, the calibration of the eye tracker was performed.

The task was performed by showing each participant the videos in the order of the counter-balanced scheme, according to a randomized ID. After each video, the participants summarized the content of the scene in 2–3 sentences. The participants were free to decide when to start the next video, in order to rest between the tasks.

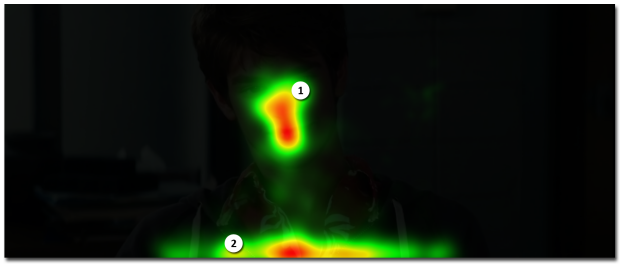
After the videos, a questionnaire was handed out to capture subjective impressions of the participants. The participants rated the layouts on a six-point Likert scale with respect to different aspects. Details about the questions will be provided in the corresponding result section. The experiment took about 60 min and each participant was compensated with EUR 10.

RESULTS

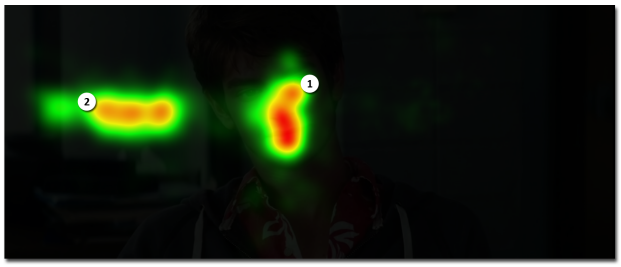
We present the results of our study in two ways: statistical evaluation and visual analysis. For statistical evaluation, we present descriptive and inferential statistics for hypothesis testing. To further support our results, we present visualizations of the data that emphasize the spatio-temporal distribution of the data with a space-time cube.

Statistical Evaluation

To test the above hypotheses (**H1–H6**) we apply inferential statistics to the following dependent measures: fixation count, fixation duration, and saccade length. Additionally, we defined AOIs for the subtitles and faces in all videos, which were used to measure fixation counts on the AOIs and transition counts between AOIs. After testing for normal distribution with the Shapiro-Wilk test, we applied either a t-test (in the case of normal distribution) or the Mann-Whitney U-test (99% confidence interval). All inferential statistics were calculated with *SPSS*. Figure 4 and Table 2 summarize the results.



(a) Regular subtitle



(b) Speaker-following subtitle

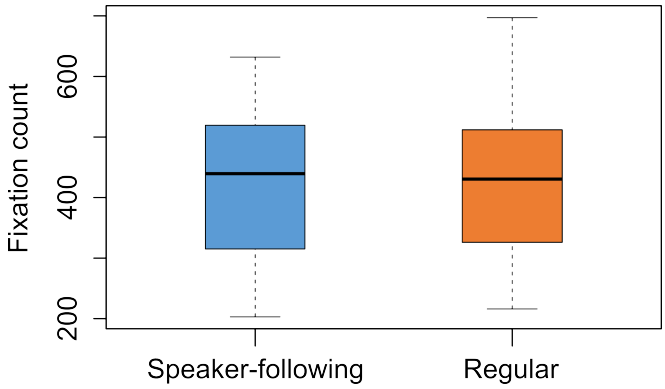
Figure 3: Attention map of approximately 10 seconds of video. (a) Regular subtitles show many gaze points on a horizontal line at the bottom (2); the face (1) is investigated occasionally. (b) Speaker-following subtitles show fewer gazes on subtitles.

Visual Analysis

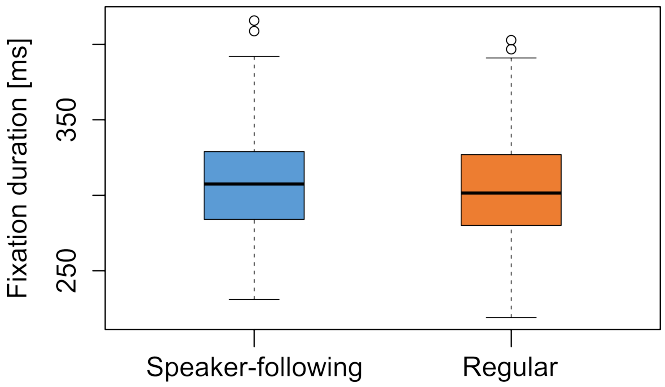
For the visual analysis of the recorded eye-tracking data, we used *ISeeCube* [24]. This visual analytics system provides support for the display of gaze data from multiple participants with attention maps and in a space-time cube (STC). Attention maps show an aggregated representation of the data. In attention maps, spatio-temporal changes, which are especially important in videos, can only be investigated by animation. In contrast, the STC visualization shows this spatio-temporal distribution of attention in a static representation, which lends itself to a better visual analysis of the connection between space and time.

The attention maps (Figures 3a, 3b) show the distribution of attention over approximately ten seconds (the first shot) of one of the stimuli. The visualizations already demonstrate that the distribution of attention differs between the layouts. Both attention maps show hotspots in the center of the screen (1), where the face of the actor was in the shot. The second hotspot (2) depends on the position of the subtitles. In Figure 3a, the second important region is at the center-bottom. In Figure 3b, the algorithm placed the speech bubble left to the face at the same height.

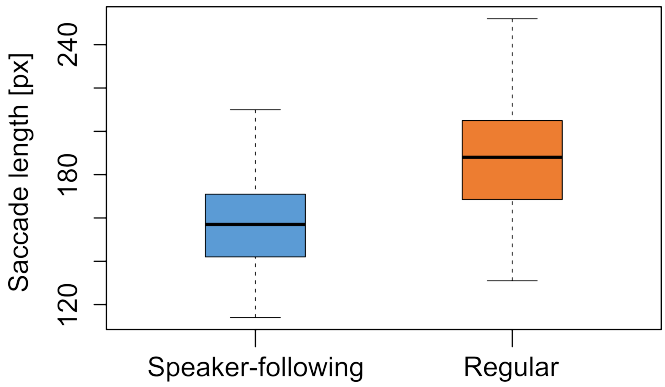
Figures 5a and 5e show the STCs of the same video with the two subtitle layouts. The dimensions of the STC are the x- and y- coordinate of the stimulus and an additional t-axis for time. The gray planes show y-t and x-t projections of the data. The gaze points are plotted with color coding: red indicates that they are close to their centroid at that time, which allows us



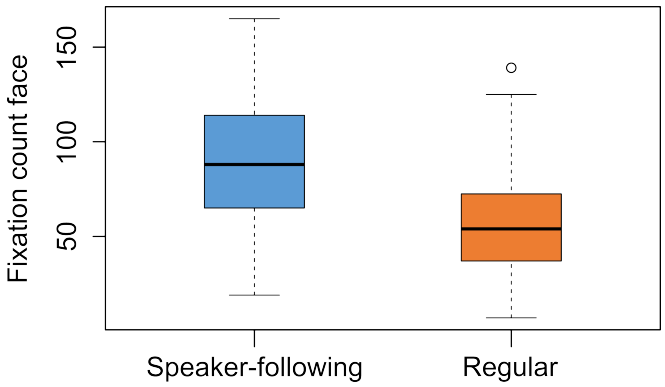
(a) Average fixation count: speaker-following (median = 439.5, mean = 423.7, sd = 111.9), regular (median = 430.5, mean = 421.9, sd = 107.9). No significant difference according to U-test (U = 19737, N = 200, p = 0.82), **H1** not confirmed.



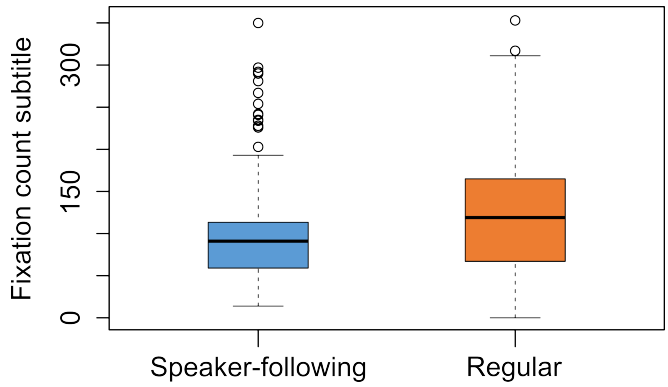
(b) Average fixation duration (in ms): speaker-following (median = 307.5, mean = 308.1, sd = 34.2), regular (median = 301.5, mean = 304.5, sd = 35.3). No significant difference according to t-test (t(398) = 1.05, p = 0.29), **H2** not confirmed.



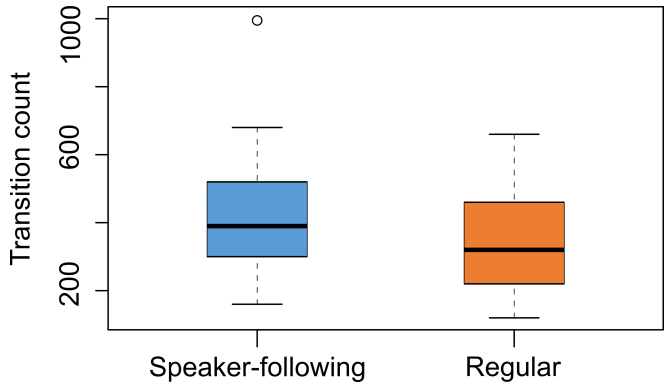
(c) Average saccade length (in pixels): speaker-following (median = 157.0, mean = 156.1, sd = 18.5), regular (median = 188.0, mean = 187.4, sd = 24.9). Significant difference according to t-test (t(398) = -14.3, p < 0.01), **H3** supported.



(d) Average fixation count on faces: speaker-following (median = 88.0, mean = 88.6, sd = 31.9), regular (median = 54.0, mean = 58.0, sd = 26.3). Significant difference according to U-test (U = 9269, N = 200, p < 0.01), **H4** supported.

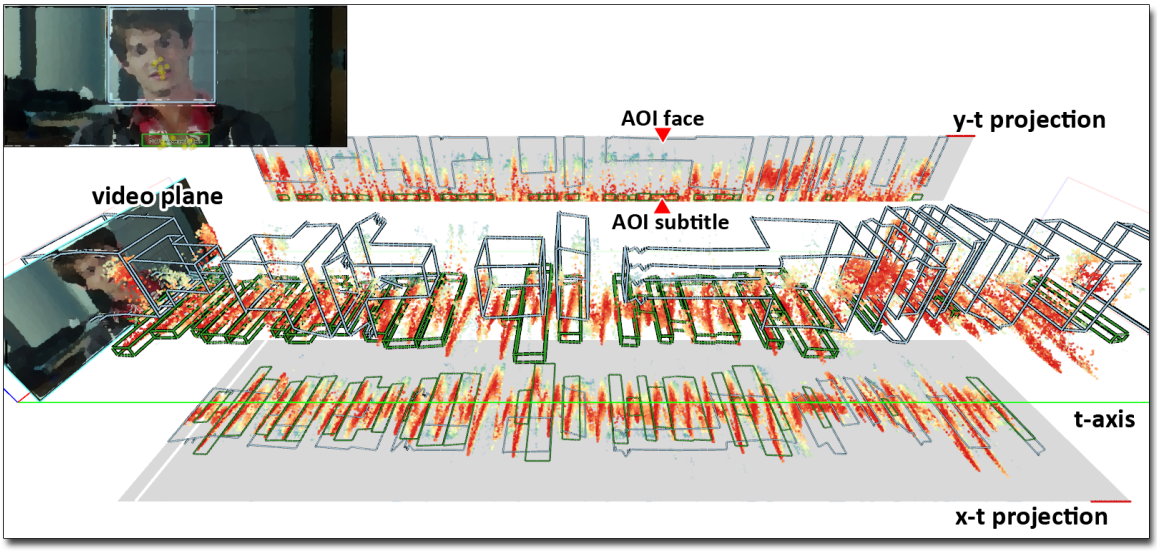


(e) Average fixation count on subtitles: speaker-following (median = 91.0, mean = 97.5, sd = 58.7), regular (median = 119.0, mean = 122.7, sd = 73.1). Significant difference according to U-test (U = 15143, N = 200, p < 0.01), **H5** supported.

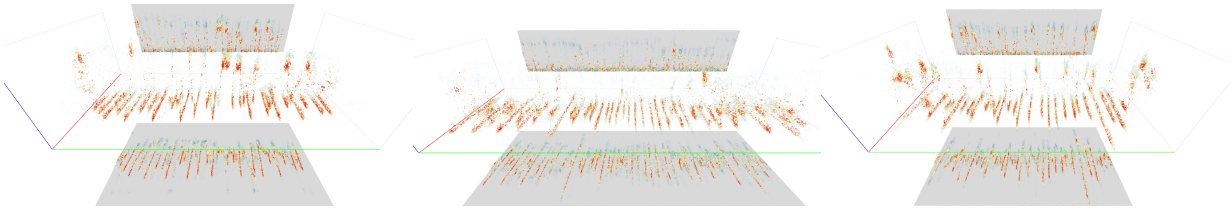


(f) Average transition count per video between subtitles and faces: speaker-following (median = 388.0, mean = 457.1, sd = 242.3), regular (median = 315.5, mean = 350.0, sd = 172.5). No significant difference according to t-test (t(18) = 1.139, p = 0.27), **H6** not confirmed.

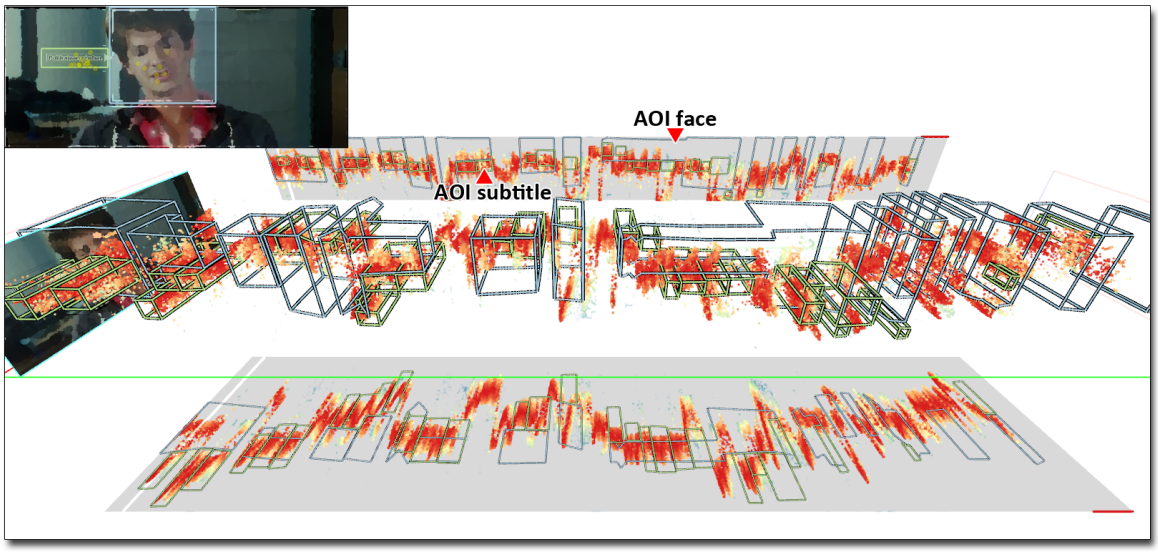
Figure 4: Boxplots of objective measures derived from the recorded eye-tracking data. Whiskers represent the lowest / highest values within the 1.5 interquartile range of the lower / upper quartile. Figures (a)–(c) are derived from the data directly, (d)–(f) are measures based on annotated areas of interest.



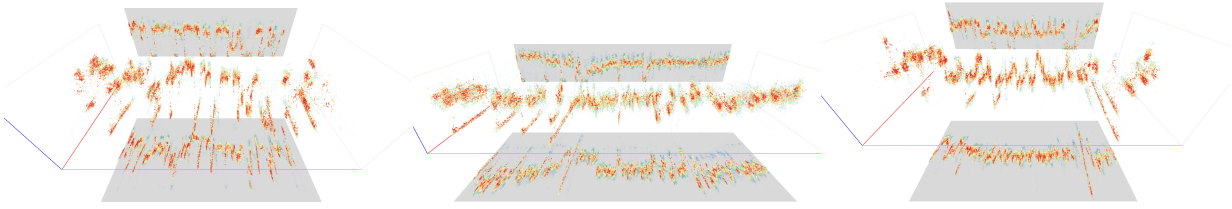
(a) STC – regular subtitle



(b) (c) (d)



(e) STC – speaker-following subtitle



(f) (g) (h)

Figure 5: Space-time cube (STC) visualization of the distribution of attention over the complete video. For regular subtitles (a–d), horizontal reading patterns on the bottom become clearly visible over time. The speaker-following subtitles (e–h) lead to a distribution that is more focused on the image content.

Table 3: Overview of the questionnaire. A (*) in the last column indicates a significant difference according to the Wilcoxon test.

Aspect	Question	Scale	Wilcoxon ($p < 0.01$)
Effort	How much effort was needed to solve the task?	(1) few–(6) much	
Visibility	How much did the subtitle impair your view on the video?	(1) few–(6) much	
Content	How well could you follow the events in the video?	(1) bad–(6) good	(Z = -3.10, N = 40)*
Readability	The subtitles were easy to read.	(1) disagree–(6) agree	
Search	I had to search for the subtitles before I could read them.	(1) disagree–(6) agree	(Z = -4.28, N = 40)*

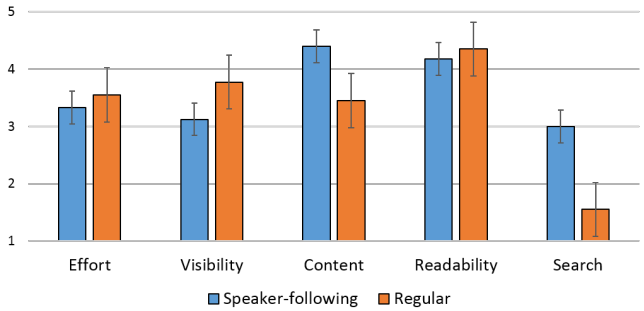


Figure 6: Questionnaire results for subjective measure of the participants' impressions with corresponding error bars.

to highlight attentional synchrony between participants. AOIs can be displayed as wire-frame shapes that also show changes in position and size of an AOI and also when it is visible during the video. Contrary to regular attention maps, we can investigate the whole timespan without animation. Figure 5a clearly shows that the focus on the center-bottom is not only restricted to the timespan depicted in the attention map. In the y-t projection, we can even see that the attention on the face seems to decrease over time and most participants are mainly reading the text at the bottom. This pattern is also visible in the other examples (Figures 5b–5d) for regular subtitles.

Figure 5e shows the spatial proximity of face and subtitle AOIs. More points are in the face AOI. In contrast to regular subtitles (a–d), the horizontal reading pattern is not as prominent (f–h). In general, the resulting pattern is closer to natural viewing behavior without subtitles. Example of general patterns without subtitles can be found in Kurzhals et al. [25]. These findings also support the quantitative results of Brown et al. [7].

Questionnaire Results

The subjective rating of the different subtitle layouts was measured on a six-point Likert scale. The Wilcoxon signed-rank test was applied to test for significant differences between the two layouts. Table 3 summarizes the questionnaire, Figure 6 lists the results. We identified a significant difference for the aspects *content* and *search*. The participants could subjectively follow the video better with speaker-following subtitles, yet at the cost that they had to search more for the subtitles before they could read them.

Finally, the participants were asked which layout they would prefer to watch movies. 19 participants decided for the speaker-following subtitles, 21 for the regular subtitles. According to additional comments from the participants, most of them were excited about the alternative layout in the beginning. Those who decided against the speaker-following subtitles in the end preferred a static position where the subtitles will appear. For speaker-following subtitles, an additional search for the next subtitle is required and some of the participants therefore preferred what they were used to.

DISCUSSION

The results of our study (Table 2) support the intention of speaker-following subtitles: the attention of the participants is less distracted from the image content while subtitles are still as readable as the regular ones.

Contrary to our findings from the pilot study, our final results cannot support the hypotheses **H1** and **H2**. Considering the number of fixations and the duration, no significant difference between the two subtitles layouts was found. In accordance with the AOI-based measures, these results indicate that only the distribution of attention changed, the fixation count and duration did not. Changes in line breaks and the resulting change in the aspect ratio could still have an influence on the resulting viewing behavior. Further investigations are required to inspect this aspect individually.

With significant differences in saccade length between the layouts, we can support hypothesis **H3**. The viewing angle between subtitles and important image content is decreased with the speaker-following subtitles. Figure 5 shows the difference of the resulting distribution of attention. Also, the subjective impression of the participants was that they could investigate the content better with speaker-following subtitles.

Considering the fixation counts on AOIs showing subtitles and faces, we could find significant differences that support the hypotheses **H4** and **H5**. The speaker-following subtitles changed the distribution of attention in favor of the image content. This means that participants spent more attention to the scene, as they would if they were watching a movie with regular subtitles. Their subjective impressions also reflect that fact. Although less attention was spent on the subtitles, the participants did not have the impression that the readability was impaired in the alternative layout.

The transition count for both layouts showed no significant differences. Although a slight increase in the transition count can be noticed (Figure 4f), we cannot confirm hypothesis **H6**. The layout has no significant influence on how often the participants switch between the text and the speaker.

Brown et al. [7] applied a regular spatio-temporal grid to calculate the differences in viewing behavior between subtitle

layouts and video content alone. They could successfully quantify the stronger deviation of viewing behavior when regular subtitles are watched. However, this approach is limited to an overall measure of gaze distribution with which it is hard to differentiate between semantic regions of the stimulus. Defining AOIs is an established method to include such semantic interpretation. Given the assumption that the selected stimuli were professionally edited to guide the viewers' attention to the appearing persons, we define AOIs based on faces and subtitles. This approach allows further interpretation of the changes in gaze behavior presented in this paper. It should be mentioned that this approach does not investigate gaze behavior on regions outside the AOIs. Although our general assumption that participants mainly focus on the subtitles and faces is true, subtle motion (e.g., hand weaving) can also attract attention, influencing the overall gaze distribution.

In summary, we can support three of our six hypotheses. The most important advantage of the speaker-following subtitles is the reduction of the saccade length. The participants' subjective impressions also confirm that they could better follow the content with less distraction by the subtitles. As a result, more fixations were on important scene content and less on subtitles, without an influence on the readability.

One limitation of our study is the fact that it focuses only on two of the three factors important for watching multimedia content. These factors are the image content, the subtitles, and the audio content [9]. For a controlled evaluation of the visual components, the videos were presented without the audio track. However, sound (e.g., the voice of the speaking person) can also guide visual attention [13]. In our study, we assume that the participants' rating for search effort of the subtitles was partially influenced by the lack of audio. If the video has an audio track, even if the spoken language is unknown, it is easier to identify the currently speaking person and therefore the position of the following subtitle. Note that our results show that even with this additional search effort, more attention will be spent on the relevant image content. The search effort was also the main reason for participants to prefer the regular subtitles. If this initial search for new subtitles can be reduced (e.g., by audio), the speaker-following subtitles could be a promising alternative to regular subtitles, that guide attention closer to the scene. Given the demographic sample of participants, our results apply to participants of ages 18 to 33. It is possible that with increasing age, the viewing behavior could differ from the presented results. Further experiments will be necessary to inspect this factor.

CONCLUSION

We conducted a user study in which we compared regular subtitles with speaker-following subtitles. Our results show that context-sensitive text placement in movies and TV shows helps keep the viewer's attention closer to the image content.

To this point, our study results relate to eye-tracking data recorded from short video clips under controlled conditions. For future work, we want to extend our studies to unrestricted environments (e.g., the living room) and full-length movie content. Further studies that also include hearing-impaired participants could provide more insight into the influence of audio on the search behavior of speaker-following subtitles. To further improve this search aspect, additional techniques to guide attention, such as visual saliency modulation [34], could be applied. Future studies should also investigate how the viewing behavior changes when text labels close to an AOI are represented in interactive scenarios such as in virtual and mixed reality [27]. An alternative approach to reduce the search effort could be the application of eye tracking as an input device. Instead of pre-calculating individual speech bubbles, positions could be determined by an attentive display [3, 14] or with a predictive approach [4], taking the viewer's gaze position into account for the layout calculation. A similar, although not interactive, method has been presented recently by Akahori et al. [1]. A comparative study between the different layout strategies could provide further insights into how an interactive approach would influence gaze behavior.

In general, the concept of speaker-following subtitles can be applied to all multimedia content that includes a speaker the subtitles can relate to (e.g., news on public displays, virtual avatar interactions). With viewers accustomed to the alternative subtitle layout, we expect speaker-following subtitles to reduce the viewing effort in the long term.

ACKNOWLEDGMENTS

We thank the German Research Foundation (DFG) for financial support within project B01 of SFB/Transregio 161.

REFERENCES

1. W. Akahori, T. Hirai, S. Kawamura, and S. Morishima. 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*, pages 123–128, 2016.

2. A. T. Bahill and L. Stark. Overlapping saccades and glissades are produced by fatigue in the saccadic eye movement system. *Experimental Neurology*, 48(1):95–106, 1975.

3. P. Baudisch, D. DeCarlo, A. T. Duchowski, and W. S. Geisler. Focusing on the essential: Considering attention in display design. *Communications of the ACM*, 46(3):60–66, 2003.

4. R. Bednarik, H. Vrzakova, and M. Hradis. What do you want to do next: A novel approach for intent prediction in gaze-based interaction. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 83–90, 2012.

5. L. Bergen, T. Grimes, and D. Potter. How attention partitions itself during simultaneous message presentations. *Human Communication Research*, 31(3):311–336, 2005.

6. M.-J. Bisson, W. J. Van Heuven, K. Conklin, and R. J. Tunney. Processing of native and foreign language subtitles in films: An eye tracking study. *Applied Psycholinguistics*, 35(2):399–418, 2014.

7. A. Brown, R. Jones, M. Crabb, J. Sandford, M. Brooks, M. Armstrong, and C. Jay. Dynamic subtitles: The user experience. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, pages 103–112, 2015.
8. D. Chiaro, C. Heiss, and C. Bucaria, editors. *Between Text and Image: Updating Research in Screen Translation*, volume 78. John Benjamins Publishing, 2008.
9. G. d'Ydewalle and W. De Bruycker. Eye movements of children and adults while reading television subtitles. *European Psychologist*, 12(3):196–205, 2007.
10. G. d'Ydewalle, C. Praet, K. Verfaillie, and J. Van Rensbergen. Watching subtitled television automatic reading behavior. *Communication Research*, 18(5):650–666, 1991.
11. G. d'Ydewalle, J. Van Rensbergen, and J. Pollet. Reading a message when the same message is available auditorily in another language: The case of subtitling. In J. O'Regan and A. Lévi-Schoen, editors, *Eye Movements: From Psychology to Cognition*, pages 313–321. Elsevier Science Publishers, 1987.
12. P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6):381–391, 1954.
13. T. Foulsham and L. A. Sanderson. Look who's talking? sound changes gaze behaviour in a dynamic social scene. *Visual Cognition*, 21(7):922–944, 2013.
14. D. Holman, R. Vertegaal, C. Sohn, and D. Cheng. Attentive display: Paintings as attentive user interfaces. In *CHI Extended Abstracts on Human Factors in Computing Systems*, pages 1127–1130, 2004.
15. K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer. *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press, 2011.
16. R. Hong, M. Wang, M. Xu, S. Yan, and T.-S. Chua. Dynamic captioning: Video accessibility enhancement for hearing impairment. In *Proceedings of the ACM International Conference on Multimedia*, pages 421–430, 2010.
17. Y. Hu, J. Kautz, Y. Yu, and W. Wang. Speaker-following video subtitles. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 11(2):32:1–32:17, 2015.
18. F. Karamitroglou. A proposed set of subtitling standards in Europe. *Translation Journal*, 2(2):1–15, 1998.
19. C. M. Koolstra, A. L. Peeters, and H. Spinhof. The pros and cons of dubbing and subtitling. *European Journal of Communication*, 17(3):325–354, 2002.
20. B. Kothari, J. Takeda, A. Joshi, and A. Pandey. Same language subtitling: a butterfly for literacy? *International Journal of Lifelong Education*, 21(1):55–66, 2002.
21. I. Krejtz, A. Szarkowska, and K. Krejtz. The effects of shot changes on eye movements in subtitling. *Journal of Eye Movement Research*, 6(5):1–12, 2013.
22. J.-L. Kruger, E. Hefer, and G. Matthew. Measuring the impact of subtitles on cognitive load: Eye tracking and dynamic audiovisual texts. In *Proceedings of the Conference on Eye Tracking South Africa*, pages 62–66, 2013.
23. J.-L. Kruger and F. Steyn. Subtitles and eye tracking: Reading and performance. *Reading Research Quarterly*, 49(1):105–120, 2014.
24. K. Kurzhals, F. Heimerl, and D. Weiskopf. ISeeCube: Visual analysis of gaze data for video. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 43–50, 2014.
25. K. Kurzhals and D. Weiskopf. Space-time visual analytics of eye-tracking data for dynamic stimuli. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2129–2138, 2013.
26. D. Miniotas. Application of Fitts' law to eye gaze interaction. In *CHI Extended Abstracts on Human Factors in Computing Systems*, pages 339–340, 2000.
27. P. Mohr, B. Kerbl, M. Donoser, D. Schmalstieg, and D. Kalkofen. Retargeting technical documentation to augmented reality. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 3337–3346, 2015.
28. E. Perego, F. Del Missier, M. Porta, and M. Mosconi. The cognitive effectiveness of subtitle processing. *Media Psychology*, 13(3):243–272, 2010.
29. A. Poole and L. Ball. Eye tracking in human-computer interaction and usability research: Current status and future prospects. In R. . D. Hyona, editor, *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, pages 573–605. Elsevier Science, 2003.
30. D. J. Rajendran, A. T. Duchowski, P. Orero, J. Martínez, and P. Romero-Fresco. Effects of text chunking on subtitling: A quantitative and qualitative examination. *Perspectives*, 21(1):5–21, 2013.
31. K. Rayner. The perceptual span and peripheral cues in reading. *Cognitive Psychology*, 7(1):65–81, 1975.
32. N. M. Ross and E. Kowler. Eye movements while viewing narrated, captioned, and silent videos. *Journal of Vision*, 13(4):1–1, 2013.
33. A. Szarkowska, I. Krejtz, Z. Klyszejko, and A. Wieczorek. Verbatim, standard, or edited?: Reading patterns of different captioning styles among deaf, hard of hearing, and hearing viewers. *American Annals of the Deaf*, 156(4):363–378, 2011.
34. E. E. Veas, E. Mendez, S. K. Feiner, and D. Schmalstieg. Directing attention and influencing memory with visual saliency modulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1471–1480, 2011.