Towards the usage of pauses in audio-described videos

Benoît Encelle
Université de Lyon,
CNRS Université Lyon 1, LIRIS,
UMR5205, F-69622, France
+33 472 431 636
bencelle@liris.cnrs.fr

Magali Ollagnier Beldame
Université d'Avignon et des Pays du Vaucluse,
CNRS, Centre Norbert Elias,
UMR8562, France
+33 490 162 736
Magali.Ollagnier-Beldame@univ-avignon.fr

Yannick Prié
Université de Nantes,
CNRS Université Nantes, LINA,
UMR6241, Nantes, France
+33 240 683 248
yannick.prie@univ-nantes.fr

ABSTRACT
Classical audiodescription process for improving video accessibility sometimes finds its limits. Depending on the video, required descriptions can be omitted because these may not fit in the durations of "gaps" in the video soundtrack (i.e. "void" spaces between dialogues or important sound elements). To address this issue, we present an exploratory work that focuses on the usage of "artificial" pauses in audio-described videos. Such pauses occur during the playing of the video so as to transmit more audio-descriptions. Our results show artificial pauses offer a good acceptability level as well as a low disturbing effect.

Categories and Subject Descriptors
H.5.2 [**User Interfaces**]: Evaluation/methodology; K.4.2 [**Social Issues**]: Assistive technologies for persons with disabilities.

General Terms
Design, Experimentation, Human Factors

Keywords
Video accessibility, video enrichment, pauses in audiodescription.

1. INTRODUCTION
The amount of video available on the Web is continually growing and as a result, video content appears as a first-choice medium to share information. However, while efforts have been made to improve the global accessibility of Web pages (e.g. Web Accessibility Initiative efforts - www.w3.org/WAI), videos still suffer from a lack of accessibility solutions and challenge a lot of accessibility problems for visually impaired/blind people [1].

The ACAV project (Collaborative Annotation for Video Accessibility) [2][3][4] addresses these issues. Our approach is based on video annotations rendered as video enrichments during the playing of the video stream. In this article we present an exploratory work that focuses on the usage of "artificial" pauses in videos enriched with audio-descriptions. Such pauses occur during the playing of the video so as to transmit audio-descriptions that require a transmission time higher than "gaps" duration in the video soundtrack. Gaps are "void" spaces between dialogues or important sound elements that can be filled without harming too much the understanding of the video.

Our results show artificial pauses offer a good acceptability level as well as a low disturbing effect.

W4A2013 – Communication, May 13–15, 2013, Rio de Janeiro, Brazil. Co-Located with the 22nd International World Wide Web Conference.

Section 2 presents related work on possible enrichments for improving video accessibility to visually impaired/blind people. Section 3 describes questions we address to assess the relevance of our artificial pauses in audio-described videos. Section 4 focuses on the experiments we conducted with blind people in order to determine how such pauses in audio-described videos can be relevant. We finally discuss our work and highlight some future work in sections 5 and 6.

2. RELATED WORK
Enriched videos are videos augmented with various elements, such as captions, images, audio assets, hyperlinks, etc. Video accessibility relies on enrichments, as they are used to translate parts of video content so that people who cannot fully understand it visually or aurally can apprehend it.

Focusing on enrichments dedicated to visually impaired/blind people, these are mostly audio enrichments that describe key visual elements, added to the sound track of the video. Different kinds of audio enrichments can be considered: pre-recorded audio files, vocal synthesis, or audio notifications (auditory icons and earcons [5]).

Innovative works have recently focused on audio enrichment. For instance, the E-inclusion project [6][7][8] took benefits of audio enrichments based on vocal synthesis. It aimed to assist humans in generating and rendering video description for blind or visually impaired people. The E-inclusion prototype uses computer-vision technologies to automatically extract visual content, associate textual descriptions to segments and add them to the audio track with a synthetic voice. The ACAV project [2] went one step further. In particular, we investigated the use of speech synthesis and earcons for enhancing the understanding of videos [4]. These projects also highlight the fact that audio-descriptions should be personalized to end-users [7] and suggest personalization mechanisms [3][8], in contrary to the "one size fits all" used by the classical audiodescription technique.

The audiodescription technique is a more classical means to provide access to theatre, television and film for visually impaired/blind people. Video audio-descriptions consist in recorded text pronounced by actors, aligned with gaps in the original soundtrack of the video (i.e. each of these gaps is associated with an audio-description). Numerous standards exist for audiodescription [9][10].

However audiodescription has limits. Each audio-description has to fit in the associated gap of the original soundtrack whatever the amount of key visual content: choices have to be made with regards to the balance between content and available gaps, sometimes resulting in the loss of useful descriptions. It then appears that classical audiodescription finds a limit regarding the amount of transmitted descriptions. To address this problem, [4] suggests to take advantage of parallel communication offered by multimodality to improve the quantity of descriptions. Other innovative ways of enriching video can also be investigated.

3. EXTENDING GAPS WITH PAUSES

We propose to perform "artificial" pauses during the playing of a video in order to transmit audio-descriptions that require a transmission time higher than the soundtrack's gap durations. We suggest placing these pauses just after gaps. As a result, for a given gap followed by a pause, the associated audio-description is transmitted during the gap and goes on during the pause duration.

Our general questioning is related to the potential usefulness of these pauses for enriching the audio-description experience for blind people. More precisely, we focus on the following questions: are pauses containing audio-descriptions perceived by blind people? Can these pauses be more useful than disturbing and under what conditions? (e.g. Have durations of pauses got an importance: is there a threshold?)

4. EXPERIMENTATION

4.1 Goals and Hypothesis

We carried out an experimentation using a mixed approach, combining quantitative and qualitative analysis methods. To study the relevance of artificial pauses in audio-described videos, we hypothesized that:

H1: The longer the pause duration is, the more disturbing it is – (*pause duration variable* V1). The goal is to measure a potential pause duration threshold beyond which viewer felt discomfort; viewer's "illusion of continuity" [11] could be broken.

H2: The higher the pause ranking number is, the less disturbing it is – (pause location variable V2). Pauses are ranked according to their start time positions. The pause that has the lowest start time (i.e. the closest to the beginning of the video) has the rank number 1; the next one has the rank number 2 and so one. The goal is to measure a potential appropriation effect of the artificial pause enrichment.

H3: Pauses are more disturbing during the first viewing of an enriched video rather than during its second viewing – (*viewing number variable* V3). The goal is to measure a potential appropriation effect of the material (i.e. a given enriched video).

4.2 Experimental material

A nearly unknown cartoon entitled "Tong" (8 min 59 sec long) was chosen. The first objective was to identify the key visual elements of this video that had to be described during some gaps, whatever gap durations. 41 key visual elements were established and shortly described according to audiodescription guidelines [10]. 41 textual short descriptions, each one associated to a particular gap, were written.

Secondly, the video was enriched so as to have an audiodescription version of the video, thanks to the Advene plateform (advene.org). We firstly annotated the video with annotations of type "*description*". Each of these annotations corresponds to a description and its associated gap: the annotation is time-aligned with the beginning of the gap and its content consists in the textual description. The audiodescription version is thus produced by speech-synthetizing all these annotations during the playing of the video. In this way, each speech-synthetized annotation corresponds to a particular enrichment (i.e. a particular audio-description).

Thirdly, after playing this enriched video, we identified 26 audio-descriptions that did not fit in associated gaps because these required a transmission time higher than gaps duration. As a result, 26 annotations of type "*artificial pause*" were added. Each of these annotations was time located at the end of its corresponding gap. Durations of these pauses have now to be clarified.

From our point of view, the duration of a gap in the soundtrack often conveys a meaning: we chose to have durations of pauses proportional to durations of gaps. More precisely, according to V1, different pause durations have to be tested. We decided to have three kinds of pause durations: D1) pause duration is equal to 0,5 * the duration of the gap, D2) pause duration is equal to the duration of the gap and D3) pause duration is equal to 1,5 * the duration of the gap.

In order to independently test V1 and V2, three enriched versions of the video were designed to have each pause, depending on the version, assigned to a duration of either kind D1, D2 or D3. For each enriched version, we did alternate pauses of kind D1, D2 and D3 using a Latin square process (e.g. D1, D2, D3, D2, D3, D1, D3, D1, D2, etc.) in order to avoid any experimental bias due to subjects' habituation effect.

For each enriched video version, descriptions associated to pauses of kind D2 and D3 were "extended" to fit pause durations. As a result, pauses of kind D1 were associated to their corresponding short initial descriptions. Pauses of kind D2 were associated to medium descriptions (i.e. a more verbose version of corresponding short descriptions). Pauses of kind D3 were associated to long descriptions (i.e. a more verbose version of corresponding medium descriptions). All the experimental material is available at http://liris.cnrs.fr/~bencelle/w4a13

4.3 Protocol and Experimental Conditions

Subjects were divided into 3 groups, each one corresponding to a specific enriched version of the video.

Subjects were in front of a personal computer containing enriched versions of the video in order to avoid potential streaming delays. Before starting, we introduced to subjects the notion of artificial pause and read the experimental instructions "*We ask you to watch the enriched video twice. During these two viewings, if you feel discomfort with an artificial pause and associated description, please press the keyboard key having an embossed sticky label. Please permanently keep a finger on it".* After watching twice the enriched version corresponding to their group, subjects filled in a questionnaire that we designed so as to collect their feedbacks.

According to our experiment hypothesis, three independent variables were studied. *V1*: the *pause duration variable* with 3 possible values: D1, D2 or D3. *V2*: the *pause location variable*: 26 possible values (26 artificial pauses). *V3*: the *viewing number variable*: 2 possible values (first viewing or second viewing of an enriched version of the video).

The number of artificial pauses perceived as uncomfortable was collected for 1) each pause duration (V1), 2) each pause location (V2) and 3) each viewing number (V3).

4.4 Participants and data collection

18 unpaid legally blind volunteers (12-69 years old, 9 males and 9 females) were recruited thanks to an association for blind people and a school dedicated to blind students. All were blind diagnosed with a visual acuity less than 1/50 after correction and a luminance perception or visual field less than 5 degrees. All subjects had no hearing disabilities.

Firstly, quantitative data about locations of discomforts felt by a blind subject during the viewing of an enriched version of the video was recorded and a tool was used to save in a specific file the time codes corresponding to key pressures.

Secondly, quantitative and qualitative data was collected with a questionnaire. The questionnaire was made up of 10 questions validated by three reviewers. Most of questions used Likert satisfaction scales: 5 closed questions concerned the global experiment (e.g. "this experiment has interested me"), 4 others closed questions dealt more precisely with the effects of the artificial pause enrichments (e.g. "pauses helped me to understand the story" / "pauses disturbed me") and 1 opened question let the subject give its impressions regarding a) this new kind of enrichment and b) its potential utility in terms of video understanding.

4.5 Results

4.5.1 Quantitative oriented analysis

For each hypothesis and associated variables (V1, V2 and V3) a Chi-square test was established to check variable possible relation with the perception of discomforts.

H1: The longer the pause duration is, the most disturbing it is

Table 1 first data line shows the observed number of discomforts according to the kind of pause duration (V1). Pauses that are of duration of type D3 seem to cause a bit more discomfort than D1 or D2. There is almost no difference concerning D1 and D2.

Chi-square test

Table 1. Observed number of discomforts / pause durations

Number of discomforts / pause durations				
	D1	**D2**	**D3**	**Total**
discomfort	22	21	26	69
not discomfort	290	291	286	867
Total	312	312	312	936

The following values are obtained: *chi-sq*=0,6571, *df*=2 and *p-value* = 0,71995 > 0,05. This statistical test emphasizes the fact there is no dependency between the two variables. Pause duration (V1) (according to our definition) seems to have no impact on the perception of discomforts.

H2: The higher the pause ranking number is, the less disturbing it is

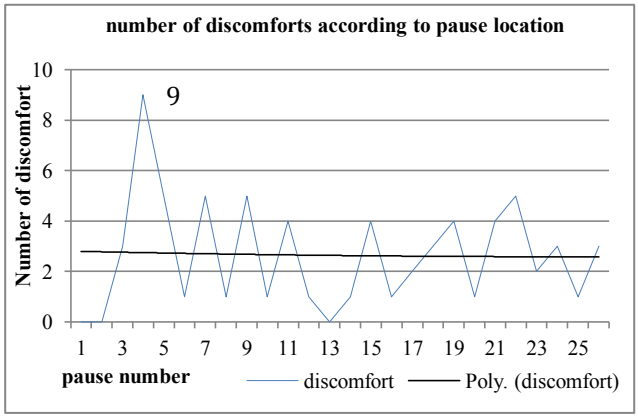

Figure 1. Number of discomforts / pause location

Figure 1 shows the number of discomforts according to locations of pauses (V2). There is a maximum of discomforts (9) for the 4st pause.

Chi-square test

The following values are obtained: *chi-sq*= 46,3283, *df*=25 and *p-value* = 0,00588 < 0,05. As a result, this statistical test emphasizes the fact there is a dependency between the two variables. Pause location (V2) seems to have an impact on the perception of discomforts.

H3: Pauses are more disturbing during the first viewing of an enriched video rather than during the second viewing

Table 2. Observed number of discomforts / viewing number

	1st Viewing	**2nd Viewing**	**Total**
Discomfort	45	24	69
not discomfort	423	444	867
Total	468	468	936

Table 2 shows the number of discomforts according to the viewing number (V3).

Chi-square test

The following values are obtained: *chi-sq*= 6,9000, *df*=1 and *p-value* = 0,00862 < 0,05. This statistical test emphasizes the fact that there is a dependency between the two variables. Viewing number (V3) of the enriched version of the video seems to have an impact on the perception of discomforts.

4.5.2 Qualitative oriented analysis

The experiment

Table 3 shows for each question concerning the experiment and for each satisfaction category the number of subjects.

Table 3. Questions about the experiment

Question	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
pleasant experiment	10	8	0	0	0
pleasant story	9	6	2	0	1
self-motivated	13	5	0	0	0
incomprehensible story	2	1	1	3	11
annoying experiment	0	0	0	3	15

According to table 3, subjects were all self-motivated to do the experiment. Generally speaking, the story was well understood and pleased the subjects. As a consequence, discomforts indicated by subjects and collected during the viewings, are not due to an annoying video or experimentation: this tends to confirm results of the quantitative analysis.

Effects of the artificial pauses

Our objective is a) to know whether or not artificial pauses were perceived and b) to estimate pauses potential utility concerning the improvement of video understanding.

According to table 4, pauses seem to be perceived and are judged more helpful than disturbing. The speech synthesis we used was clearly qualified as understandable.

Overall utility of artificial pauses in enriched videos

According to results we collected from the opened question, pauses seem to be helpful when gap durations are too short to transmit needed descriptions of key visual elements. For instance, some subjects emphasized that this kind of enrichment could all the more be useful for audio-describing videos that have intense rhythm (e.g. action movies). One subject argued that pauses could also be useful to convey visual-only information, for instance text-on-screen information and time changes (i.e. flashback, etc.). In order to be less surprising, some subjects suggested more

integration between pauses and the video original soundtrack. Their proposal consisted in doing video soundtrack fades-out/fades-in respectively before/after pauses. Two subjects stated that pauses generally broke the story as they were expecting the rest, and as such, felt discomfort. One subject stated that pauses were not uncomfortable but not useful either. Two subjects stated that pauses find their usefulness when users can trigger them. For instance, pauses can be triggered in order to repeat a misunderstood audio description or when a user wants to generally slow down the speech synthesis speed.

Table 4. Effects of the artificial pauses

Question	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
Pauses perceived	9	7	0	1	1
Pauses helpful	6	5	4	1	2
Clear synthesized voice	10	4	2	2	0
Pauses disturb	2	3	5	4	4

5. DISCUSSION

Concerning the variable *pause duration* V1, according to our quantitative analysis, the kind of duration D3 collected a bit more discomfort than D1 or D2. No statistical relation could be established between chosen kind of durations and the notion of discomfort. As a result, we did not find a "threshold" concerning pause duration that clearly separates our pause durations into groups (i.e. subjects accepted pause durations/ rejected (discomfort) pause durations). However, in order to deeply inspect the possible existence of this threshold, additional experiments with a bigger difference between pause durations have to be conducted.

Concerning the variable *pause location* V2, we hypothesized that the lower the pause ranking number is, the more disturbing it is. According to the results (cf. Figure 1), the maximum of discomforts is not associated to the first pause (i.e. the closest pause from the beginning of the video), but corresponds to the 4th pause (nearly close to the beginning). This may be due to a short experiment adaptation delay from our subjects. Indeed, they were facing a new kind of video enrichment and may have required an adaptation time, a training period before being ready for the experiments. However, as one of our goals consisted in measuring a potential appropriation effect of artificial pause enrichment (V2), we preferred to bypass this training period. According to the quantitative analysis, the pause location is in relation with the perception of a discomfort and according to the trend curve (cf. Figure 1) we could notice that the level of discomfort seems to be slowly decreasing along the time.

The variable *viewing number* (V3) seems to be in relation with subject perception of discomfort. This variable was set up in order to check if there is an appropriation effect of the material - the enriched video – by the subjects. Indeed, the number of discomforts strongly decreases between two viewings. As a result, the end-users acceptability of artificial pauses in videos seems to be good. This interpretation is reinforced by the qualitative subject answers that in majority are satisfied by this new kind of enrichment.

6. CONCLUSION / FUTURE WORK

In this paper we present an exploratory work that focuses on the usage of "artificial" pauses in audio-described videos. These pauses occur during the playing of the video so as to transmit more audio-descriptions. Our results show artificial pauses offer a good acceptability level as well as a low disturbing effect. This exploratory work is a first investigation of the usage of pauses in audio-described videos. Another key modality in video enrichment dedicated to the blind, to be studied in the future, will be Braille display.

7. ACKNOWLEDGMENTS

This paper was supported by the French Ministry of Industry (*Innovative Web call*) under contract 09.2.93.0966, "Collaborative Annotation for Video Accessibility" (ACAV) and by the ARC (Academic Research Communities) program of the Rhône-Alpes Region, France.

Authors warmly thank Nasthasia Kovacs for her implication on presented experiments, analysis and results. We thank Olivier Aubert, the main engineer of the Advene. We are indebted to the experiment participants; to Cité Scolaire René Pellet in Villeurbanne and Association Valentin Haüy (AVH) in Lyon for providing local arrangements.

8. REFERENCES

[1] Burgstahler, S. Creating video and multimedia products that are accessible to people with sensory impairments.
DO-IT, University of Washington
http://www.washington.edu/doit/Brochures/PDF/vid_sensory.pdf. Accessed 02/15/13.

[2] Champin, P.-A. and Encelle, B. et al. 2010. Towards collaborative annotation for video accessibility. In *Proc. W4A '10*. 17:1–17:4.

[3] Saray Villamizar, J. F. and Encelle, B. et al. 2011. An adaptive videos enrichment system based on decision trees for people with sensory disabilities. In *Proc. W4A '11*. 7:1–7:4.

[4] Encelle, B. and Ollagnier-Beldame, M. et al. 2011. Annotation-based video enrichment for blind people: a pilot study on the use of earcons and speech synthesis. In *Proc. ASSETS '11*. 123–130.

[5] Blattner, M. and Sumikawa, D. et al. Earcons and Icons: Their Structure and Common Design Principles. *Human➔Computer Interaction*, *4*(1), 11–44.

[6] Gagnon, L. and Foucher, S. et al. Towards computer-vision software tools to increase production and accessibility of video description for people with vision loss. *Universal Access in the Information Society*, *8*(3), 199-218.

[7] Chapdelaine, C. 2010. In-situ study of blind individuals listening to audio-visual contents. In *Proc. ASSETS '10*. 59–66.

[8] Chapdelaine, C. and Gagnon, L. 2009. Accessible videodescription On-Demand. In *Proc. Assets '09*. 221–222.

[9] ITC 2000. ITC Guidance on standards for audio description. Technical report ITC.

[10] Morisset, L. and Gonant, F. 2008. Charte de l'audiodescription. Technical report .

[11] Berliner, T. and Cohen, D. J. The Illusion of Continuity: Active Perception and the Classical Editing System. *Journal of Film and Video*, *63*(1), 44–63.