AMSE JOURNALS-AMSE IFRATH Publication –2014-Series: Modelling C; Vol. 75
Special Handicap; pp 95-105
Submitted Sept. 2014; Accepted Oct. 31, 2014

Video Enrichment for Sensory Impaired People: Design Issues and Pilot Studies on Perception

* Magali Ollagnier-Beldame, ** Benoit Encelle, ***Yannick Prié

* CNRS, Centre Norbert Elias, UMR8562, University of Avignon and Pays du Vaucluse,
France, Avignon, (Magali.Ollagnier-Beldame@univ-avignon.fr)
** CNRS University Lyon 1, LIRIS, UMR5205, University of Lyon
France, Lyon F-69622, (bencelle@liris.cnrs.fr)
*** CNRS University of Nantes, LINA, UMR6241, University of Nantes,
France, Nantes, (yannick.prie@univ-nantes.fr)

Abstract

Enriched videos are videos that are enriched with various elements during their play, such as captions, images, audio, hyperlinks, etc.: such enrichments could improve video accessibility to sensory impaired people. We firstly define precisely what video enrichment is. Secondly, we present two pilot studies related to perception of two kinds of audio enrichments for videos. More precisely, participants of these studies had to interact with variously enriched videos containing metaphorical auditory icons and speech synthesis enrichments. These studies indicate different trails to follow in future experiments, related to quality, variability and quantity of enrichments.

Key words: Video enrichment, perception, multimodality, video accessibility

1. Introduction

Video is now more tightly integrated into hypermedia documents on the web (*e.g.* video embedding in web pages, fragment based access to video from hypertext links, navigation links within video, *etc.*). Amongst the various possibilities offered by such video / hypermedia convergence, we are interested here in *enriched videos*, namely videos that are enriched with various elements during their play. If such enrichment is not new, as evidenced by the long history of subtitling, the possibility of considering *annotations* as independent temporal information that will be rendered while playing the video stream, offers a wide range of

variability on how videos can be enriched. There is indeed a wealth of variation in the contexts of use of enriched videos; in the categories of annotations that can be considered and in the ways the rendering can be done. Among these possibilities, we focused on several kinds of video enrichments that can improve video accessibility for sensory impaired people. As a main contribution, we conducted two pilot studies about perception of two kinds of audio enrichments: auditory icons and speech synthesis enrichments. Our various findings show that video enrichment perception study is indeed a valuable research field and they point out some trails to consider for future experiments.

The article is organized as follows: in section 2 we define video enrichment and its uses. Section 3 deals with general principles for producing enriched videos, based on annotations. We then focus on the question of video enrichment perception for sensory impaired people, presenting related works (section 4) followed by two pilot studies (section 5) that deal with innovative kinds of video enrichments for blind or visually impaired people, before concluding.

2. Video enrichment for translating or complementing a video

We define video enrichment as consisting in *enriching the playing of a given video with supplementary elements* such as textual descriptions, images, audio or video fragments, hyperlinks, *etc.* Two main goals can be followed while creating an enriched video: *translating* part of its content so that people that cannot fully understand its visual or audio content can apprehend it; or *complementing* it with additional information so as to enhance the watching experience. As sensory impaired people benefit from translation-based video enrichment, we focused here on this kind of enrichments.

Traditionally, for deaf and hearing-impaired users, *closed-captions* can be used to describe the audio track of a video (dialogs, sounds, music). For visually impaired people, *audio description* consists in adding verbal information to the audio track in order to describe the visual content of the video [1].

3. Annotation-based video enrichment

Considering the technological / computer science point of view on video enrichment, the most fruitful approach seems to consider *annotation-based video enrichment,* both for analyzing existing systems or designing new systems. A video annotation is here defined as *any information associated to a fragment of a video*. For instance, a text describing a part of a dialog can be associated to a temporal fragment (defined by two timecodes) during which this part of dialog

occur, while a text describing part of a scenery (*e.g.* “notice the knife”) can be associated to a spatio-temporal fragment [2].

Annotation data can be *rendered* so as to enrich a video. Annotation rendering consists in presenting the content of an annotation in an adequate modality, *e.g.* visually enriching the video track with textual captions, drawings, still images, video fragments, *etc.* or auditory enriching the audio track with voices, music or sounds. Tactile modality is now being considered for instance with vibrating devices or Braille displays. The content of annotations can be rendered without a change in modality (*e.g.* by captioning a textual content or playing a sound content) or with a change (*e.g.* by speech synthesizing a textual content or using an image associated with it, for instance the photo associated to an actor’s name). Annotating is the activity of producing annotations. It consists in 1/ identifying pertinent fragments in the video, and 2/ adding information to them. Annotating is usually done using annotations tools. Such process can be performed in an automatic way by analyzing the video and/or the audio signal(s) (*e.g.* for shot segmentation or speaker diarisation, speech recognition) or in a semi-automatic way when human validation is needed. It can also be manual, when human expertise and interpretation is needed (*e.g.* describing the bad position of one hand in a piano lesson) or for more creative annotations (*e.g.* funny comments on a scene).

Rendering tools are often based on engines that follow rendering rules to enrich a video with a set of annotations. For instance, if we consider a subrip (SRT) file as a set of annotations associated to a video, most video players can enrich it by adding subtitles, following a single simple rule such as: *IF the current timecode of the player is X AND there is an annotation that has Y as its beginning timecode and Z as its ending timecode) AND* $Y \leq X \leq Z$ *THEN render the text of the annotation at the bottom of the screen*.

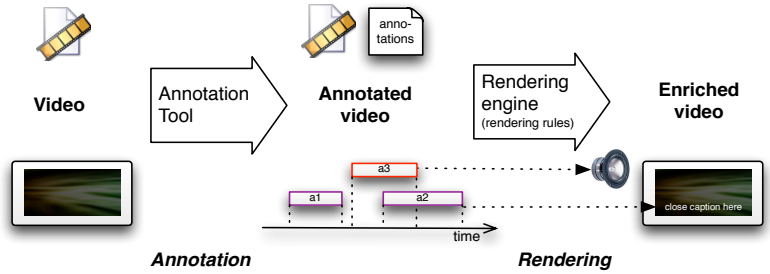

Fig. 1. The general process of annotation-based video enrichment, from annotating to rendering. Annotations a1 and a2 are rendered as closed captions, while a3 is rendered aurally.

The separation of the video (content to be enriched) from the annotations (content used for enrichment) as well as the rendering rules (how to enrich) has good properties: it can for instance foster innovation by allowing different people to create annotations and rendering independently.

It also allows to perform “live” video enrichment according to end-user preferences that can change during the rendering itself, paving the way to real-time adapted enrichments.

4. Related works on perception and enrichments for sensory impaired people

Amongst the whole scope of situations that implicate apprehension of enriched videos, situations related to sensory disabilities are "extreme” ones with regards to the effects of multimodal integration. In this context, the addition of information to videos clearly has the objective of facilitating or even to allow its perception and understanding. For example, for visually impaired people, [3] studies the combination of haptic stimulation with Braille, combining a standard Braille cell with a force feedback device. This situation corresponds to a single sensory channel (tactile) with several types of information (verbal and non-verbal). The results of his study indicate that performance in reading and comfort are worst than with the use of a simple Braille device. [4] propose a system that provides multimodal feedback by offering auditory information (non-verbal sounds and voice synthesis) reinforced by tactile information. The system leads to reading performances and comfort inferior to the reading on a standard Braille display. [5] study “semiotic” translation of information with “graph sonification”, using auditory icons to make graphs accessible to the visually impaired. Preliminary evaluation of their work indicates that both visually impaired and sighted people can understand the patterns of graphs by listening to auditory graphs. [6] evaluate multimodal perceptive systems for blind people. His results show that it is not efficient to replace all the information supplied by vision with a unique modality (for example audio) because it would imply too strong concentration needs for blind users. [7] and [8] also show that the use of a single non-visual modality for blind people give less good results in the task of recognizing the objects of a graphical interface than the use of combined modalities [9]. [10] propose a prototype system designed to enrich deaf people’s experience of music by using non-audio input information channels. Their system has two main components: a vibrating “Haptic Chair” and a computer displaying visual effects corresponding to some features of the music. Their results show that the haptic chair has the potential to significantly enhance the musical experience of a deaf person. These results show that some associations of modalities for presenting pieces of information seem to work better than others. They also show that these associations can be thought either in terms of semiotic modalities or in terms of sensory modalities. Finally, they bring to light the essential role of appropriation and learning of "new" associations for enriched videos.

In the following section, we expose two experiments for evaluating subject's perception of two kinds of video enrichment. These experiments follow the "concordance of information" principle : videos are enriched with consistent information according to their content.

5. Two experiments to assess enriched videos perception

The first experiment focused on the perception of video enriched with auditory icons [11], the second of a combination of auditory icons and vocal synthesis.

5.1 First experiment: auditory icons enrichment

Goals and Hypothesis. The aim of the experiment was to assess the perception of auditory icons as enrichments and the role of these icons in the understanding of both the editing of the video and the rhythm of the video story. The hypothesis was that the understanding of the enriched video would be better when the number of icon categories used increased. We thus expected that, according to the increase of the complexity of experimental conditions, results would be close to those of a control population.

Protocol. The subjects were 16 blind French people from 16 to 45 years old. A control population of 17 French sighted people between 22 and 29 years old also participated to a pretest about the understanding of the story. The idea of the experiment was to enrich the audio with metaphoric auditory icons at every shot change and for every camera movement. For instance, when a shot change was a "cut", a sound of cutting scissors was added and for a "travelling" camera movement, we used the sound of a wagon on rails.

Three levels of complexity related to the perception of the editing where identified: 1/ distinguishing changes, being shot transitions or camera movements; 2/ distinguishing between simple shot transitions vs. sophisticated ones (*e.g.* cut *vs.* fading) and between movements of the body of the camera vs. movements of its lens; 3/ apprehending the precise type of every shot transition or camera movement. We then designed nine experimental conditions for video enrichment, with three general levels of complexity combining various complexity levels of shot transitions and camera movements.

We used a 4 minutes and 20 seconds excerpt of a cartoon movie, which, according to the conditions and the level of associated complexity, included from 26 to 66 auditory icons. Subjects had headphones.

An enriched video was presented to each subject (phase 1) that next answered a questionnaire (phase 2). The questionnaire contained 11 questions. The two first questions dealt with the perception (or the non-perception) of auditory icons and on the possible associated

embarrassment. The 9 remaining questions dealt with the understanding of the rhythm of each scene, with possible answers ranging on a six-point bipolar scale from a slow to fast rhythm. Phase 3 consisted in a one-hour focus group that was conducted with the participants, mainly on the quantitative and qualitative properties of the presented auditory icons.

Results. The questionnaire results showed that all the subjects had heard some icons. But the results from the focus group showed that for many subjects, *all* the icons were not heard. This seems to be bound to their nature: we had indeed chosen icons that had a meaningful metaphorical link with what they denoted. However, such a choice turned out not to be relevant because the discrimination of icons with regard to the audio track was prevented.

The quantitative results showed that our hypothesis could not be validated. Indeed, with the increase of the number of icon categories, the understanding of the rhythm of the movie decreased instead of increasing. The cognitive load of the subjects was probably too important, all the more so since icons were delivered without documentation. Finally, our experimental proposition did not succeed in transmitting rhythm elements of the movie to blind persons.

Conclusion. This experiment showed that adding auditory icons is a promiseful way to enrich videos, but the number of icon categories (regarding learning) and their quality (regarding their discrimination against the original audio track) must be carefully taken into account.

5.2 Second experiment: combination of auditory icons and speech synthesis

Goals and Hypothesis. The aim of this experiment was to study two types of audio enrichments (auditory icons and speech synthesis) for blind people, added to a movie. Particularly, our objective was to estimate the role of auditory icons in blind people's understanding of the settings of the scenes and the role of speech synthesis in the understanding of the actions of the characters. Our first hypothesis was that the more auditory icons categories there are, the less the understanding of the settings is good. This first hypothesis aimed at detecting a possible threshold in the number of icons categories beyond which they do not help the understanding of the movie any more. Our second hypothesis made the assumption that a simplified content for the speech synthesis will result in a better understanding of the actions of the characters.

Protocol. This experiment was composed of three phases: a) enriched videos were presented to blind subjects, b) Subjects answered a questionnaire concerning their understanding and perception of the movie and c) they participated to a focus group.

Concerning the experimental material for the first phase, two short (less than 4 minutes each), self-sufficient (no movie excerpt) videos were enriched with auditory icons (to describe

setting changes) and with speech synthesis (to describe the actions of the characters). The first video (V1) featured two settings and two characters; the second (V2) featured six settings and a single character.

The auditory icon enrichments varied as follow for Video 1 (2 settings):

- AI1 (Auditory Icon 1): use of single icon (always the same) to indicate that the setting changes. During the first presentation of the icon, a vocal message follows the icon and announces its content: "change of scene".
- AI2: use of two different icons, each associated with one setting. At the first presentation of each icon, a vocal message follows the icon and announces its content: “in the lounge" or " in front of the scene".

For Video 2 (6 settings):

- AI1: same as for video 1.
- AI2: two different icons for two subcategories of settings. Two vocal messages are used after the first presentations of the icons: “in the kitchen” or “in a bathroom”.
- AI4: four different icons for four categories of settings. Four vocal messages: "at her window", "in her bathroom", "in her toilets" or "in her kitchen".
- AI6: six different icons for six settings. Six vocal messages: “at her window”, "in her bathroom", "at the bathtub", "at the washing machine", "in her toilet", "in her kitchen".

The descriptions of the actions of the characters with speech synthesis varied as follows. For every video we had two conditions:

- sSS (simplified Speech Synthesis): simplified description of the actions, for example “Lying down, she reads”.
- dSS (detailed Speech Synthesis): detailed description of the actions, for example “Lying down with a pillow under the head, she reads a comics”.

All in all we thus had several conditions of enrichment with auditory icons and speech synthesis for every video:

- Video 1: 4 conditions (AI1_sSS, AI2_sSS, AI1_dSS, AI2_dSS).
- Video 2: 8 conditions (AI1_sSS, AI2_sSS, AI4_sSS, AI6_sSS, AI1_dSS, AI2_dSS, AI4_dSS, AI6_dSS)

The subjects were young blind persons from 13 to 21 years-old, all fluent in Braille. We had 8 subjects: 2 for each condition of video 1 and one for each condition of video 2. We balanced the attribution of enriched movies according to the level of speech synthesis enrichment: Each subject had a video enriched with sSS and a video with dSS, for example: V1 (AI2_dSS) and V2 (AI1_sSS).

The subjects were individually seated in front of a computer, with headphones. They viewed each enriched video twice (phase 1) and then answered a questionnaire (phase 2). The questionnaire contained 21 questions for each video: 1 question about the perception of auditory icons (yes/no), 5 questions about settings (4 multiple-choices questions and 1 free answering), 5 questions about the actions of the characters (multiple-choices) and 10 questions about the appreciation of the audio track. Those questions proposed a bipolar scaling method: five-point Likert scale, measuring either positive or negative response to a statement. Phase 3 consisted in a one-hour and a half focus group. The discussion mainly focused on the utility of the sound that had been added to the audio track and the "pleasant" or "unpleasant" aspect of this audio track.

Results. Firstly, results concerning our experimental material (essentially obtained in the focus group) show that the quality of auditory icons and vocal synthesis was very good. However, the quantity of speech synthesis (i.e. the number of words added to the movie) was too important and led to overlaps with the dialogues in the video. Redundancies between the texts of speech synthesis and the audio track of the movie have been pointed.

Secondly, the question about the perception of auditory icons led to the following results: for both movies, when a single icon was present, all the subjects heard the icon. When there were 2 icons, 87 % of the subjects heard them. When there were 4 or 6 icons in the enriched video, only 50% of the subjects heard them. These results show either an effect of the number of icons categories (more than 4 different icons would be too much) or of their quality (problem with discriminating icons from the movie audio track).

The question on the number of settings in the movies (question with free answer), had theses results: we obtained 75 % of right answers for V1 (100 % for AI2 and only 50 % for AI1). For V2, the answers were mostly false, the closest answer (5 settings instead of 6) was obtained with a single auditory icon (AI1). When evaluating the frequency of the changes in the settings, 81% of the subjects declared that these changes occur very often or often. The questions about the presence of characters in settings had 81% of correct answers for V1. 60% of the incorrect answers happened in single icon conditions. For V2 there was 78% correct answers, equally distributed amongst the four AI-related conditions.

Concerning the actions of the characters (accessible by speech synthesis), we obtained 84% of good answers for V1, and incorrect answers are not correlated to sSS or dSS conditions. For V2 we obtained 80% of right answers, and 70% of false answers occur within long description condition.

From these first results, we can say that our first hypothesis relative to the negative effect of the number of auditory icons on the understanding of settings is not validated. Indeed, even if for

V2 a light effect appeared, we also had strong effects in the other direction for V1: it is in situation where the subjects have two different icons instead of a single one that the best answers are given. Our second hypothesis concerned the positive effect of speech synthesis with simplified content on the understanding of the actions of the characters. This hypothesis is partially validated with V2 because the detailed speech synthesis was at the origin of the great majority of the errors.

Thirdly, the answers related to the appreciation of the audio tracks gave these results: for both videos, the audio track pleased 60% of the persons (33 % of the participants did not answer the question). 62% found that the enriched audio track was not difficult to understand while 25% felt a difficulty (12% did not answer the question). 93% of the participants found that they rapidly became used to the sounds that add been added to the audio track (only 5% did not answer). 43 % of the persons answered that the addition of auditory icons does not require additional documentation (apart from the announce at the first time the icon appears), while 25% expressed their need for such documentation to be actually able to understand icons (30% did not answer). 81% of the subjects found that the audio track is not overloaded (18 % did not answer). 75% estimated that the icon-enriched audio track helped them understand the story, while 12% did not agree with this statement and 12% did not answer. 87% considered that understanding the audio track did not raise problem, 6% underlined it was problematic (6% did not answer). 62% of the subjects liked the audio track while 25 % did not like it (12% did not answer). 93% estimated that the first listening of the audio track did not raise any problem (6% did not answer). Finally 93% of the persons assessed that the audio track is not boring while 6 % found it boring.

As a conclusion to these appreciation questions, we can see that the enriched audio track was well received by the participants. In particular, the subjects became rapidly used to auditory icons, even from the first listening, and without being bored by the enriched audio track. A great majority of the subjects found that the understanding of the audio track was easy and that auditory icons helped them to understand the story. The audio track mainly pleased the participants. On the other hand, it also emerges from these answers that the main weakness of our proposal was the lack of documentation associated with the presentation of auditory icons (55% of the subjects were in favour of a documentation or did not answer this question).

Fourthly, the discussions in the focus group emphasized that auditory icons constituted the main innovative point (with regards to more classical enrichments), and it has been judged as very promising. The main objective of the use of auditory icons is to decrease the use of speech synthesis that "encroaches" inevitably on the original audio track. It also appeared that the use of auditory icons would be especially relevant for temporal (flashbacks, projection in the future) and

spatial or spatiotemporal (dreams, imaginary worlds) movements in the movies. The "time of the movie" was also quoted as an element that would benefit from auditory icons: icons would be used for example for the dividing into chapters or to signify the presence of *black screens*.

However the criterion of the number of auditory icons seemed fundamental to the usability of this kind of information, and their possible control by the speech synthesis was evoked so as to increase their appropriation, the icons coming then to complete or to modulate a message presented using speech synthesis. But the main problem raised in the focus group about auditory icons concerns their learning. 50% of the participants spontaneously explained that speech synthesis, besides being widely used, asked for less important concentration than auditory icons, in particular because it has the same "type" as the main element of movies: the dialogues. This remark is completely in agreement with what we know about the treatment of the various channels of a movie and about the separation between the verbal and the non-verbal rather than between the sensory modalities - *cf.* Dual Coding Theory [12].

Conclusion. This experiment mainly showed the strong potential of auditory icons for video enrichment, even if numerous questions remain about their actual use.

Conclusion

In this paper we firstly define precisely what enriched videos are and stress their utility in order to improve video accessibility for sensory impaired people. We thus introduce general principles, based on video annotations, for producing enriched videos. As a main contribution, we conducted two experiments with blind subjects related to the assessment of the perception of various video enrichments: auditory icons and speech synthesis. Our preliminary results tend to show that the usage of pertinent enrichments that are concordant with video contents seems to have a positive impact on their understanding. *Auditory icons* for visually impaired can be used to transmit time and space information (*e.g.* movie rhythm, settings). They have to be easily distinguishable from the audio track (be careful with metaphoric icons). For learning reason and in order to avoid cognitive overloads, the number of different icons should not be greater than 6. Moreover an auditory icons index that could be accessed any time during the playing of video could enhance the efficiency of the icons.

Finally and generally speaking, a good balance between the quality of the enrichments, their variability and their quantity has to be found. These experiments are pilot studies: further experiments have to be conducted to precise and complete these results.

References

1. L. Gagnon, S. Foucher, M. Heritier, M. Lalonde, D. Byrns, C. Chapdelaine, J. Turner, S. Mathieu, D. Laurendeau, N.-T. Nguyen, and D. Ouellet: Towards computer-vision software tools to increase production and accessibility of video description for people with vision loss. In: Universal Access in the Information Society 8, 3. 199-218, 2009.
2. D.C.A. Bulterman : Using SMIL to encode interactive, peer-level multimedia annotations. In: ACM Symposium on Document Engineering: 32-41, 2003.
3. C. Ramstein: Combining haptic and braille technologies: design issues and pilot study. In Proceedings of 2nd ACM Conference on Assistive technologies. ACM, NY, 37-44, 1996.
4. C. Ramstein, O. Martial, A. Dufresne, M. Carignan, P. Chasse and P. Mabilleau : Touching and Hearing GUIs. Design In: Issues in PC-Access System", In proceedings of ACM ASSETS96 Conference (International ACM SIGACCESS Conference on Computers and Accessibility), Vancouver, 1996.
5. S. H. Choi and B. N. Walker: Digitizer Auditory Graph: Making graphs accessible to the visually impaired. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems. Atlanta, GA, (10-15 April). 3445-3450, 2010.
6. Y. Hatwell: Images and non-visual spatial representations in the blind. In Burger, D. and Sperandio, J.-C., editors, Non-visual Human-Computer Interactions, number 228 in Colloque Inserm, 13-35, Paris. John Libbey Eurotext Ltd., 1993.
7. A. Dufresne, O. Martial and C. Ramstein: Multimodal User Interaction System for Blind and "Visually Occupied" Users: Ergonomic Evaluation of the Haptic and Auditive Dimensions. In Proceedings of Interact'95, HCI, Chapman & Hall, 163-168, 1995.
8. O. Martial : Représentation mentale et multimodalité dans l'interaction humain-ordinateur pour des utilisateurs aveugles. In: Actes du 6ème Colloque ERGO IA'98, 49-18, 1998.
9. P. Truillet : Modélisation de coopérations intermodales : application à l'interaction non-visuelle, Thèse d'Université, Toulouse III, 1999.
10. S. Nanayakkara, E. Taylor, L. Wyse and S H. Ong : An enhanced musical experience for the deaf: design and evaluation of a music display and a haptic chair. In Proceedings of the 27th international Conf. on Human factors in computing systems. ACM, NY, 337-346, 2009.
11. W. W. Gaver. Auditory icons: using sound in computer interfaces. *Human-Computer Interaction,* 2(2), 167-177, 1986.
12. J.M. Clark, A. Paivio: Dual coding theory and education. In Educational Psychology Review, 71, 64-73, 1991.