Annotation-based video enrichment for blind people: A pilot study on the use of earcons and speech synthesis

Benoît Encelle
Université de Lyon,
CNRS Université Lyon 1, LIRIS,
UMR5205, F-69622, France
benoit.encelle@liris.cnrs.fr

Magali Ollagnier-Beldame
Université de Lyon,
CNRS Université Lyon 1, LIRIS,
UMR5205, F-69622, France
mbeldame@liris.cnrs.fr

Stéphanie Pouchot
Université de Lyon,
Université Lyon 1, ELICO,
EA 4147, F-69622, France
stephanie.pouchot@univ-lyon1.fr

Yannick Prié
Université de Lyon,
CNRS Université Lyon 1, LIRIS,
UMR5205, F-69622, France
yannick.prie@liris.cnrs.fr

ABSTRACT

Our approach to address the question of online video accessibility for people with sensory disabilities is based on video annotations that are rendered as video enrichments during the playing of the video. We present an exploratory work that focuses on video accessibility for *blind people* with *audio enrichments* composed of speech synthesis and earcons (i.e. nonverbal audio messages). Our main results are that earcons can be used together with speech synthesis to enhance understanding of videos; that earcons should be accompanied with explanations; and that a potential side effect of earcons is related to video rhythm perception.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Auditory (non-speech) feedback, Evaluation/methodology; K.4.2 [**Social Issues**]: Assistive technologies for persons with disabilities.

General Terms

Design, Experimentation, Human Factors

Keywords

Video accessibility, accessibility for blind people, video annotation, video enrichment, audio notification.

1. INTRODUCTION

Accessibility to digital information for all, including people with disabilities, is one of the major social challenges of our society. Laws were voted to support this idea, e.g. “section 508” in the USA or a 2005 law in France. The United Nations also adopted the Convention on the Rights of Persons with Disabilities [31] in 2006. However, while efforts have been made to improve the accessibility of some types of electronic content (e.g. global accessibility of Web pages – Web Accessibility Initiative [33]), other types still suffer from a lack of accessibility solutions. As the amount of video available on the Web is continually growing and as its consumption is continuously increasing, video content appears as a first-choice medium to share information. In this context, the 21st Century Communications and Video Accessibility act signed in October 2010 into US law promotes expanded access to internet-based video programming. New technical solutions that allow people with sensory disabilities (hearing/visually impaired, deaf and blind people) to access this kind of content need therefore to be developed.

The ACAV project (Collaborative Annotation for Video Accessibility) addresses these problems by exploring how accessibility of online videos can be improved by developing free Web applications intended for a large audience. 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 video accessibility for *blind people* with *audio enrichments* composed of speech synthesis and earcons (i.e. nonverbal audio messages). Our main results are that earcons can be used together with speech synthesis to enhance understanding of videos; that earcons should be accompanied with explanations; and that a potential side effect of earcons is related to video rhythm perception.

Section 2 presents the scientific and technical context of the ACAV project. The related work (section 3) focuses on blind people and audio enrichments. Section 4 deals with questions concerning information access and understanding, associated with audio enriched videos. Section 5 introduces our technical proposal for enriching video with audio elements –including earcons– and focuses on the experiments we conducted with blind people in order to determine the utility and the usability of these earcons.

2. CONTEXT

2.1 Video accessibility and video enrichment

Classical techniques for improving video accessibility for people with sensory disabilities include the audio-description of key visual elements for visually impaired/blind people, and the subtitling/close-captioning and/or the sign translation of key audio elements for hearing impaired or deaf people. Deaf-blind people need a combination of these two techniques and, in some cases, with descriptions presented on a Braille display.

Concerning particular cognitive and neurological disabilities [32], some individuals may process information aurally better than by reading text: audio descriptions of text embedded in a video can be needed. For autism, the content should be customizable and well designed so as to not be overwhelming. Media adaptation has then to focus on the purpose of the content and has to provide alternative content in a clear and concise manner. Such alternative content could for instance present the key points of the video (e.g. key educational messages, important verbal communications, etc.). Another issue for autism could be to present social stories (a series of pictures, supported by simple text to describe the actions, behavior, and outcomes that some quite visual individuals might learn effectively from). A combination of pictures and synchronized text or audio could then be added to the video in order to improve its accessibility.

For all these disabilities, additional content has to be associated to the whole video or to some parts of it and presented over it using one or several modalities. We call this approach *video*

Conference'10, Month 1–2, 2010, City, State, Country.

enrichment. Its underlying concepts and the associated formats, tools and recommendations are detailed in the following section.

2.2 Enriching videos

Enriched videos are videos augmented with various elements, such as captions, images, audio, hyperlinks, *etc*. The goals are here either to *translate* parts of its content so that people who cannot fully understand it visually or aurally can apprehend it; or to *complement* it with additional information in order to enhance the watching experience. Two types of users are basically involved in a video enrichment process: users who enrich videos, namely the enrichment producers, and end-users who watch enriched videos.

Several technical recommendations, initiatives and formats related to videos enrichment have been issued recently. The Web Accessibility Initiative (WAI) advocates in one recommendation (WCAG) [34] the development of different versions of given temporal content, e.g. audio versions using audio-description of visual content, etc. Mozilla Foundation [26] advocates the usage of the Ogg open video format with multiplexed specialized tracks for video accessibility. In the same way, the HTML Accessibility Task Force suggests adding several tracks to a video content, e.g. a subtitle track, an audio-description track, etc. These "enrichment" tracks would be represented as HTML 5 *Track* elements inside a *Media* element (*Video* or *Audio* element). As an alternative to this expected notion of track in HTML 5, SMIL (Synchronized Multimedia Integration Language) can be used for synchronizing different multimedia contents (e.g. a video synchronized with an audio file containing an audio-description and with a subtitle file).

Full-featured tools for making accessible video are not yet available. Several subtitling tools nevertheless exist, such as MAGPie, Nico Nico Douga or YouTube subtitler[1]. Tools for audio-describing visual content are even less common, though MAGPie permits the recording of audio elements while playing the video.

2.3 Annotation based video enrichment

Most of the preceding formats or tools use "direct" enrichments: the added elements are presented without a change in their original modalities (*e.g*. by captioning a textual element or playing a sound element). Our approach is different in that it separates the content used for enrichment from its rendering. As a result, not only "direct" enrichments but also "indirect" enrichments (*e.g*. speech synthesizing a textual content) can be provided. In our opinion, this separation –similar to the one suggested in the document-engineering field– has good properties. It can for instance foster innovation by allowing different people independently create content or content rendering, for example in a collaborative process. It also allows performing "live" video enrichment according to end-user preferences that can change during the rendering itself, paving the way to real-time adapted enrichments.

Accordingly, the ACAV project general workflow is made up of two main steps: an annotation step and a rendering one (cf. Figure 1). The first step consists in annotating the video. An *annotation* is defined as *any information associated to a fragment of a video*.

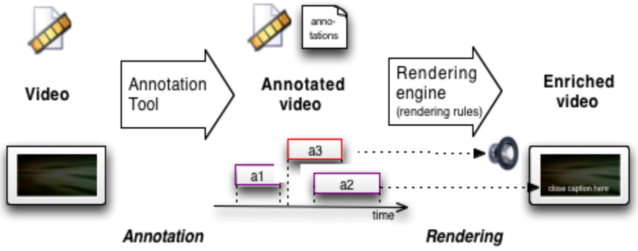

Figure 1: annotating videos for enriching their playing

For instance, a text describing an action can be associated to a temporal fragment (defined by two timecodes) during which this action occurs. The second step consists in rendering annotation data in order to enrich the video. Annotation rendering for accessibility presents the content of an annotation using one or several adequate presentation modalities. As a result, a video can be enriched using three main kinds of elements: visual enrichments (captions, still images, video fragments, *etc*.), audio enrichments (voices, sounds) or tactile enrichments (using vibrating or Braille devices).

Before we present the data models used for representing the contents of enrichments and their renderings, we first describe a general user–oriented overview of the ACAV system (Figure 2).

2.4 Towards collaborative video enrichment

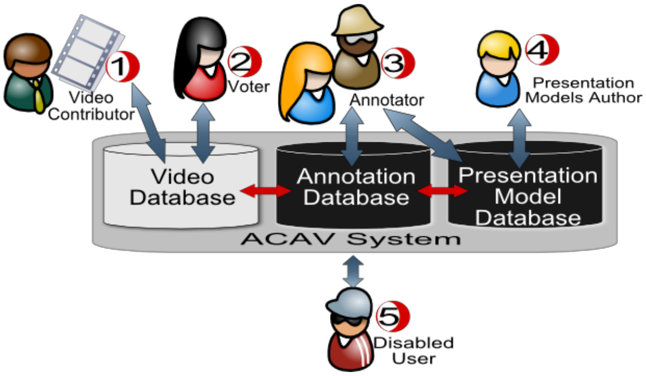

Figure 2. ACAV system and users

The ACAV system defines several workflows that were initially defined with accessibility experts and sensory disabled people. A vote system allows determining a priority order for videos that have to be made accessible (i.e. enriched) for various audience(s) (visually disabled and/or hearing disabled people). Disabled users, their friends/relatives can thus use this vote system (2). As a result, a producer of enrichments (more precisely here an annotator) can view this list of requests and start the annotation step. She also can invite other annotators (i.e. collaborators) to help her (3). After this stage, she can specify a presentation model for rendering these annotations for a given disability. She also can share her annotations with other people, for instance authors of presentation models (4).

Next, users with disabilities can query the video database and find enriched videos adapted to their interests and disabilities, view enriched videos and personalize the presentation of some kinds of enrichments (e.g. change the average rate of the voice synthesis, enlarge font sizes, etc.) (5). If an end-user encounters troubles

[1] http://ncam.wgbh.org/invent_build/web_multimedia/tools-guidelines/magpie
http://www.nicovideo.jp
http://yt-subs.appspot.com

during the visualization of an enriched video, he can use the feedback system to inform enrichment producers (back to 3/4).

2.5 Annotation and rendering models

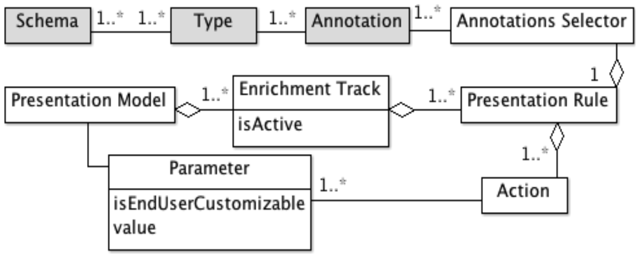

Figure 3. Main elements of annotation and rendering models

The annotation model used in ACAV (Figure 3, grey) is an adaptation of a more general model for video annotation proposed in [1]. *Annotations* are the key elements of the model. Basically, an annotation has a unique id, a content and is associated to a temporal interval. *Annotation Types* classify annotations by describing their semantics and constraining and structuring their content. Each annotation is associated to a given type (e.g. annotations of type *Character*, of type *Action*, etc.). An *Annotation Schema* embodies a particular annotation practice and is composed of several types. For example, one could define a schema for describing the dialogues of a video, another schema for the musical part, and yet another one for the shots.

The annotation-rendering model (Figure 3, white) is based on the notion of enrichment tracks (cf. section 2.2) to describe the way annotations are to be rendered. A *presentation model* contains one or several enrichment tracks. An *enrichment track* contains one or several presentation rules. A *presentation rule* is made up of two parts: an annotation selector and one or several presentation actions. An *annotation selector* is the expression of constraints that filter the annotation set. These constraints could be structural (for selecting annotations associated with one or several schemas, types) and/or intrinsic (for selecting annotations based on properties of their content). A *presentation action* indicates the modality that has to be used for rendering the contents of selected annotations. Possible presentation actions are:

- For visual disabilities-adapted renderings: *oralize* (using a text-to-speech engine), *display in Braille* (integral or contracted form), *use a Braille Symbol*, *play a sound*
- For hearing disabilities-adapted renderings: *display as a subtitle*, *display as a close-caption*, *display a shape*, etc.

A presentation action can be parameterized and related values could be, if indicated, end-user customizable. Enrichment tracks and their parameters are all that end-users can manipulate by activating or deactivating tracks or changing their parameters.

3. RELATED WORK

Having presented our general approach for video enrichment for accessibility in the ACAV project, we focus here on audio enrichment for blind and visually impaired people.

Audio-description is a means of providing access to theatre, television and film for blind or visually impaired people. For videos, the audio-description process consists in describing the visual elements of a film to give essential keys for its understanding. The recorded text is aligned with gaps in the original soundtrack of the video, and mixed with it. The fundamentals of audio-description for theater can be found in [3] and [28], while numerous standardization documents exist for video [19, 24, 27, 30]. [23] gives the core information categories that must be described for movies: *characters* (appearance/role) and their interrelations, *actions* and the *sets* used in the filming. They underline that the importance of the descriptions both depends on the type of the film and the time that is available for description.

Audio-description has its limits. First, not all the visual content is described. For instance visual editing information is generally not indicated despite the fact it can be important for understanding the story. Moreover, audio-description only uses the audio verbal modality. As a consequence, parallel communication possibilities offered by multimodal communications are unavailable. In addition, making an audio-description is money and time consuming (for instance, audio-describing a 90 min movie in France costs more than 5000 Euros (7300 USD) and takes one month). [22] also highlights the fact that audio-description should be personalized, contrary to a "one size fits all" approach, a remark that is coherent with the general approach of the ACAV project for video on the web accessibility.

Audio enrichments for the blind consist in the addition of audio information to a film sound track. Different kinds of enrichments can be added: pre-recorded audio files, vocal synthesis, or audio notifications (auditory icons and earcons). The E-inclusion project [8,22] aims to assist humans in generating and rendering video description for people who are blind or visually impaired. The E-inclusion prototype uses computer-vision technologies to automatically extract visual content, associate textual descriptions and add them to the audio track with a synthetic voice [13]. This work is different from our: there are no annotation/presentation models, no usage of different modalities (E-inclusion only uses the vocal synthesis), and no usage of audio notifications.

Audio notifications. *Auditory icons* [15, 16] are everyday sounds that convey information about events by analogy to everyday sound-producing events (e.g. the sound of crumpling paper for indicating the event "trash can empty"). Good mappings between sounds and associated meanings should therefore be easy to learn and remember. However, auditory icons *"lack flexibility, as metaphoric mappings are not always easy to find*" [14] and "*can be confused with actual environmental sounds*" [10].

Earcons are defined in [4] as "*nonverbal audio messages used in the user-computer interface to provide information to the user about some computer object, operation or interaction*". This definition is extended to "*abstract, musical tones that can be used in structured combinations to create auditory messages*", "*composed of short, rhythmic sequences of pitches with variable intensity, timbre and register*". The main advantage of earcons is their flexibility and the fact that they "*can be designed in families so that they represent hierarchies, by controlling or manipulating their different parameters (e.g. timbre and pitch)*". However, earcons suffer from a lack of meaningful relationship with their referent: end-users have to learn and memorize mappings between sounds and associated meanings.

Audio notifications have to be sparingly used as they could cause annoyance if too frequent [5]. With regards to the learning of mappings between audio notifications and associated meanings, auditory icon notifications are generally found to be easier to learn and retain in comparison with earcon notifications [4,6,7,12,17,21]. Audio notifications have been used for *conveying information that has a strong spatial component* to blind and visually impaired people. For instance, they were used for improving orientation and mobility skills of blind people (in conjunction with haptic feedback) [29], for improving objects localization [11] or for enhancing accessibility to mathematical material such as equations with fractions (two dimensional objects) [25] or graphs [9] using "graph sonification". Audio notifications have been also frequently used for *improving accessibility of Human-Computer Interface components*: e.g. audio menus and scroll bars [35], mobile service notifications [14]. However, as far as we know, audio notifications have been never used to try to convey information related to videos.

4. QUESTIONS ASSOCIATED WITH THE USE OF AUDIO NOTIFICATIONS

Several general *perception issues* are related to the use of audio information for enriching videos, with regards to information access and understanding. First, *low level perception* issues are a) the need to assess the cognitive (over)load engendered by the added sounds, and b) the need to test their discriminability, i.e. the difference from the original soundtrack (important parameters for discrimination are volume, sound duration, etc.). These issues are all the more important when added audio information pertains to various semiotic modalities [2], such as audio linguistic (speech synthesis) or audio non linguistic (audio notifications). This is the case in our research where we hypothesize that using *bi-modal* video enrichments associating earcons with speech synthesis is possible.

Second, one important question we have to face concerns *high level perception*: *how to reach integrated perception* for the users? Audio enrichment of video needs to ensure that a) audio notifications help the understanding of the video, while at the same time, b) every enrichment (including audio notifications) is smoothly integrated into the whole so as to reach a "unity of the video support". Jaskanen [20] refers to the tension between these aims with the term 'paradox'. In our context, this tension can only be clarified by qualitative experiments.

Finally, we need to question *earcons' utility* for film enrichment, in relation to the kind of video information they are useful for (Characters? Actions? Sets?). Earcons are used in operating systems to represent specific and important events, such as the end of an operation or ruptures in a temporal stream (the arrival of a mail or chat message, an error, etc.). For video enrichment, we hypothesize that, and will try to assess whether, they are useful to transmit spatial and temporal information (here we will focus on set changes indicated by earcons).

5. EARCONS AND SPEECH SYNTHESIS FOR AUDIO ENRICHMENT

We carried out several experiments using a mixed approach, combining quantitative and qualitative methods. Using an inductive methodology, we worked step by step: the conclusions from a preliminary experiment allowed us to design a second experiment and to refine our results. Section 5.1 deals with the production of our experiments material (enriched video for the blind). Sections 5.2 and 5.3 present the preliminary and the principal experiments (all the experimental material is available at: http://www.advene.org/acav/assets11).

5.1 Experimental material: enriched videos

We defined several presentation models (see section 2.5) upon one simple annotation schema in order to produce audio enriched videos for blind people. Annotation schemas indicate what to describe of the visual content of videos and rendering models specify means of presenting the resulting descriptions, as described below.

Annotation schema. Visual information is described in an annotation schema called *VisualBase* made up of the annotation types *Action* and *Set* (corresponding to the key visual information identified by [23] minus *Character*), complemented with a *TextOnScreen* annotation type to represent text appearing on the screen (e.g. opening or closing credits).

We used the *VisualBase* schema to annotate the videos we used in the tests. We used two short humorous videos V1 (3') and V2 (1'45). V1 was described by 54 annotations (average length: 8 words) describing sets, characters' actions and text on screen. V2 was described by 24 annotations (average length: 6 words). We used the following audio-description rules: annotations should 1) be brief and not tell too much, in order to preserve the original work; 2) not overlap important parts of the soundtrack; 3) be as neutral as possible. Examples of annotations are: "*in front of the stage*" (Type: *Set*), "*Ben is showing the coffee table*" and "*The girl is standing by him, is showing the table and the socks*" (Type: *Action*), "*MyBox Production presents… interpreted by… realized by…*" (Type: *TextOnScreen*).

Presentation models. Audio notifications have to be sparingly employed in order to avoid end-users annoyance (cf. section 3), and only a small number of them can be learned. Because there usually are less different sets than actions and because their succession rhythm goes generally slower, using earcon audio notifications for *Set* annotations was considered as a potentially relevant and innovative rendering.

As presented in Table 1, we produced several presentation model variations with different kinds of bi-modal audio enrichments (speech synthesis and earcons). In addition to the rendering of annotations when playing the video, we also added a prologue for each video. Figure 4 illustrates these various notions focusing on the presentation model for one particular situation.

Our goal was to study two types of enrichments: earcons and speech synthesis. More precisely, we wanted to:

- Verify that our descriptions of video content actually helped blind persons understand it.
- Identify which verbosity level is better adapted to action descriptions: simplified (C1) or detailed (C2)
- Validate the fact that the association of earcons and speech synthesis does not create too much a cognitive load that would prevent video understanding.
- Confirm that users can handle as much as 6 different earcons, a number we had estimated based on an unpublished experiment.

Participants were individually shown (on a computer) V1 and V2, both annotated with the *VisualBase* schema and audio-enriched with earcons for *Sets* and speech synthesis for other annotations (cf. PM-S0 in Table 1). As there are 2 sets in V1 (actions take place either in front of a theater stage or in a lounge), we used 2 different earcons. For the 6 different sets in V2 (6 places in one apartment), we used 6 earcons. Videos do not have the same rhythm: V1 has much more rapid rhythm than V2. After each film, participants answered 21 questions related to their perception of the audio enrichments and their understanding of the story. At the end they all participated to a focus group.

Table 1. Five presentation models for rendering annotations during the playing of a video

	PM – S0	PM – S1	PM – S2	PM – S3	PM – S4
PROLOGUE					
Video Metadata	-	Synthesize Speech for Synopsis			
Set	-	Synthesize speech for lexicon	-		
ENRICHED VIDEO RENDERING					
Action, TextOnScreen	Speech Synthesize content of the annotation				
Set	Play Earcon associated with content *then* Synthetize speech for content if first occurrence	Play Earcon associated with content	Play Earcon associated with content *then* Synthetize speech for content if first occurrence	Play Earcon associated with content *then* Synthetize speech for content if asked by the user	Play *unique* Earcon *then* Synthetize speech for content of the annotation

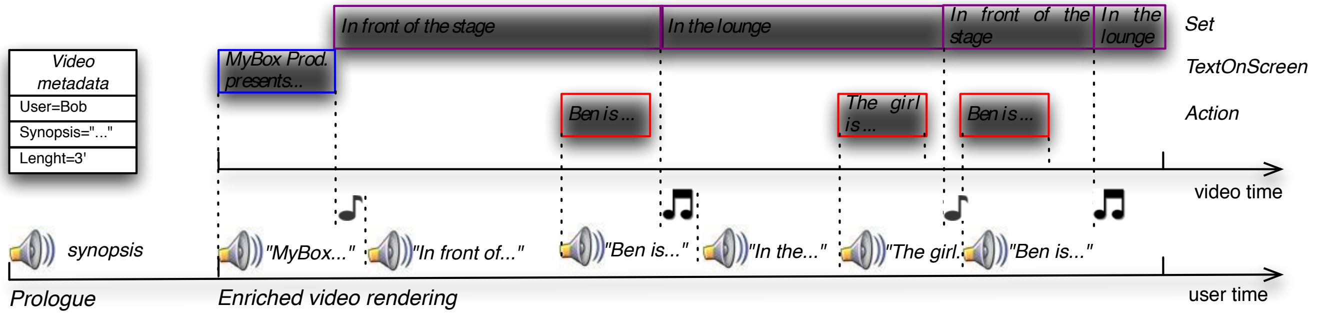

Figure 4. An illustration of audio enrichment of a video with Presentation Model S2

Our enrichment proposal received a warm welcome from the participants. Results from questionnaires showed that video understanding was very good, meaning that our annotations are relevant, but sometimes too long and sometimes overlapping the dialogues (even a little overlap was distracting).

They also showed that the better speech synthesis enrichment is the simplified one (C1), because it contains nearly all the useful information and is shorter than the detailed one. So the criteria of shortness and respect of the original soundtrack are more important than the 'exhaustiveness' one (description with a lot of details). From the focus group, the participants' comments show that earcons are well adapted to communicate time and space change information. The benefits for understanding are clear, as participants were very positive on this point: "*Hum, I found that the sounds were very interesting to mean set changes*", "*Just the small beep for the change of scene it's good.*", "*I prefer audio icons than speech synthesis* (…)*. An icon is short, while speech synthesis encroaches a lot upon the soundtrack*". As to the number of different earcons, questionnaires and focus group showed that 6 earcons could be handled. Focus group also showed that earcons are easy to learn, that their speech synthesis explanation (just after the first time an earcon is played) is useful.

These findings clearly revealed the potential of earcons for video enrichment, and showed that mixing earcons and speech synthesis could be useful. Moreover, as blind persons are used to speech synthesis and audio-description, we considered (a) that audio-enriched videos should contain audio-description, and (b) that the use of earcons alone for audio-enrichments is not conceivable yet, as a prerequisite would be societal agreement on their meaning. So we decided to focus on the way earcons and speech synthesis can be combined to express one category of description (*Sets*) while keeping "classical" audio-description for *Actions*. We then conducted a more substantial experiment with new objectives and hypotheses aimed at studying this combination in enriched videos.

5.2 Combining earcons and speech synthesis for presenting sets

Our objective here was to evaluate and compare four kinds (S1-S4) of bimodal enrichment presentation models as described in Table 1. Earcons are used each set change. Speech synthesis is used to present both the actions of the characters and the text on screen (e.g. title of the video). It is also used to pronounce some text accompanying earcons and explaining them. Here, we hypothesized that:

- H1: Earcons are perceived as additional elements to the video. This hypothesis concerns the low-level perception of earcons.
- H2: Earcons (combined with speech synthesis) help participants to understand the elements of the story they describe (here *Sets*).
- H3: Earcons help participants perceive narrative "meta-information". Earcons here indicate set changes, and should lead to high-level perception of the number of different sets and of the frequency of set changes.

Our pilot study also suggested to us that it might be useful to include a short oral overview of each video, at the beginning.

5.2.1 *Protocol and experimental conditions*

The protocol of this experiment is quite similar to that of the pilot study with some minor adjustments regarding both video enrichment and questionnaires.

For each video, a *prologue* part (cf. Table 1 and Figure 4) was added, that consisted of a short synopsis. Besides a “control” situation with no enrichment, we distinguish four “situations” (S1…S4), ways of using the earcons. In S1, S2 and S3 we used as many earcons as there are sets in the videos, with these differences:

- S1: a lexicon presenting earcons was presented before the video begins, and after the synopsis.
- S2: synthesized speech explained what each earcon corresponded to just after it was used for the first time.
- S3: the speech explanation for the last earcon heard was accessible by typing the F1 key of the keyboard.

S1, S2 and S3 are situations where speech synthesis has a “supporting” role to the earcons; speech is used here to explain sounds. But we wanted to explore a second type of combination: in S4, a unique earcon was used for every set change and a short speech synthesis was systematically associated to it as explanation (e.g. “♪/ kitchen”; “♪/ bathroom”; etc.). Earcons here show that the following descriptions concern sets and not action, they act as “semantized audio onsets” for the following speech.

These four *situations* should not be confused with more “classical” experimental *conditions* that would be set up to study the effect of variables on other variables. We aimed to investigate the possible useful earcons / speech synthesis combinations, and our quantitative results are mainly guides to assess the good ones.

Situations 1, 2, and 4 are ‘theater situations’, ones that could be used with an audience of many people. Situation 3 is more individualized: videos are showed on a personal computer, with each participant listening to the soundtrack with headphones. In this paper we present data comparing the three “theater” situations, S1, S2, and S4; S3 will be the subject of future work.

5.2.2 *Participants and data collection*

We recruited 21 unpaid legally blind volunteers (23-72 years old) with the help of an association for blind people. One of them did not watch the first video (V1). Six participants contributed to S1 (V1 and V2), eight participants to S2 (V1 and V2), six participants to S4 (V1) and seven participants to S4 (V2). All are traditional media consumers: they watch TV and listen to the radio. A few go to the movies. Some watch DVD and all agree that the use of DVD players is difficult for them and they need help to access the audio-description functions. Some of the participants have good computer skills, while others are not familiar with Internet and do not use it. For each situation, after each video, participants answered 21 questions (being helped to fill the questionnaires) about both the earcons (perception, understanding, quality of the enrichment) and their understanding of the story itself. The questions concerning the story understanding improvement thanks to earcons were ‘closed’ questions: 4 questions for V1 (Qa…Qd), 3 questions for V2 (Qe…Qg). These questions concerned ‘facts’ in the story. For example ‘Where does Ben install his piece of furniture?’. These questions thus had expected answers. At the end they all took part in a focus group that we recorded. We collected our data in the projection room of the association.

5.2.3 *Results*

Although the results we present here only concern S1, S2 and S4, we already have interesting elements. Three volunteers participated to a control condition (video without enrichments). Their answers to questionnaires were logically "reconstructed" and they said they missed information to fully understand the video, particularly concerning sets (about 50% of expected answers). They were also favorable to an earcons-based enrichment.

Concerning H1, for both videos, it appears that 85 % of the participants heard the earcons, thus confirming that earcons are readily perceptible.

More precisely, the distribution of the perceptions shows a perfect perception for S1. This leads us to further hypothesize that a preliminary presentation of icons and their meaning is effective on the future perception of earcons. Situation 2, which we considered as the “most difficult” situation, does not show a strong effect in the perception.

Our second hypothesis H2 aimed at assessing the understanding of the sets-related elements of the story. Overall, this understanding was good (Table 2). Our hypothesis is thus confirmed. However, the performance varied across questions (see discussion). Situations were grouped in this analysis, as there was no apparent effect of the situation on the understanding of the video story.

Table 2: Amount of good answers to set-related understanding questions regarding to expected answers

	Video 1 (V1)				Video 2 (V2)		
Question	*Qa*	*Qb*	*Qc*	*Qd*	*Qe*	*Qf*	*Qg*
S1, S2, S4	14/20	16/20	17/20	8/20	10/21	18/21	10/21
Total	55/80 - 69%				38/63 - 60%		

The third hypothesis H3 was related to the global understanding of ‘meta-information’ about the narration, earcons acting as indicators of the number of sets and the frequency of set changes. Regarding the evaluation of the number of sets, participants did not have the exact answer most of the time, though they were close for V1 (2 sets), and under the exact number for V2 (6 sets).

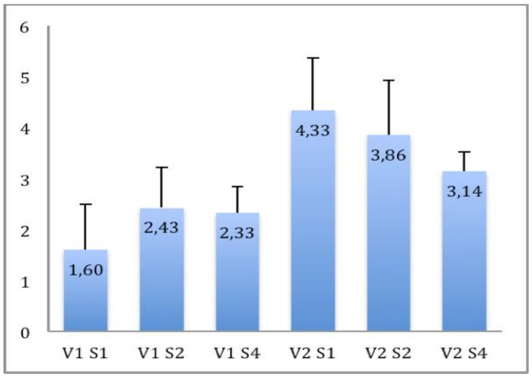

Figure 5. Means of perceived number of set changes

Figure 5 shows inter-individual agreement in S4 (where the name of the set is repeated each time), with better answers for

V1 than for V2. Regarding frequency of set changes, the results did not show much, but suggested that earcons give saliency to visual elements that where not deemed significant by non visually-impaired viewers (see discussion). Overall, our third hypothesis is not verified though there is a potential effect of the repetition of the sets' names on remembering them.

Participants also answered a questionnaire to evaluate the enriched videos they had been presented. They gave their opinion using a Likert satisfaction scale. Answers to these questions show three interesting points. First, the majority of the participants (65%) found that enrichments helped them understand videos. It is interesting to note that it is in situations 2 and 4 that the help was considered the most useful. Secondly, nearly half of the participants (46%) declared that the enriched videos had pleased them. The distribution of these answers is very uneven and the majority of S2 participants declared having appreciated the video. This is strengthened by the answers to another question which concerned boredom: all the subjects of S2 said they did not get bored, while in the other situations there were only a third. Finally, nearly half of participants (44%) found that it was easy to adapt themsleves to the enrichment, except in S1 where we had 8 negative answers of 12.

6. DISCUSSION

Our first question was about low-level perception of earcons. H1 was confirmed and the best results were obtained with S1 (preliminary presentation of earcons significations in a lexicon). Focus groups confirmed the sensitivity of blind people to the quality of earcons and their discriminability in the soundtrack. S1 obtained the best qualitative answers ("agree") to the question on easiness of perception of the enriched content. **We think the preliminary presentation of earcons has an effect on their later perception**. Focus group also brought information on how to improve the lexicon for earcons' preliminary presentation: "*we could put twice the beep then its meaning*". Another S1 participant would have preferred the lexicon was associated with a repetition of all earcons and meanings for each occurrence: "*if there are many earcons, it would be necessary to verbalize every time*". So **the lexicon must be refined and improved. In the future accessibility tool, lexicon apparition should be user customizable**.

Our second question was about high-level perception of earcons. Our third hypothesis (H3: earcons help participants perceive narrative "meta-information") was not confirmed. However it mainly showed that participants evaluated V2's rhythm as fast while we considered it medium, meaning that blind persons did not share our rhythm judgment. Using earcons for sets may lead to distorting the perception of the video rhythm, because earcons draw the attention to set changes (discontinuity). This illustrates Jaskanen paradox, leading us to state that **earcons must be perceived but they must not perturb.** A S2 participant asserted that "*it has to be non intrusive; we have to be able to make our own idea of the image*". Hence, **earcon enrichment of videos can have side effects, leading to modifications in the perception of video rhythm, possibly not desirable.** So video enrichment based on automatic video segmentation based on set classification should be used with care.

Our third question concerned the utility of earcons associated to speech synthesis. Participants' answers show a good global understanding. **Earcons associated with speech synthesis are useful for the understanding of sets-related information**. Besides we must be attentive to the contents of annotations: as the results of the pilot study show, **the criterion of conciseness dominates the criterion of exhaustiveness**. We had several comments in this direction. An S1 participant asserted: "*You should not enrich too much. Some sounds are self-sufficient and it is not necessary to describe them with speech synthesis*". Another comment from S4: "*you should not look for too much coverage, otherwise it is too heavy. It is not necessarily annoying if we do not understand locations exactly. It is not necessary to detail all the sets*".

Besides, the prologue was well received for both films in all situations and considered very useful in the focus groups. We think **that presenting a video synopsis as a prologue is beneficial for understanding**.

Our experimental results will help us specify elements of **good practices** for annotating videos and designing rendering. When the annotation platform comes on-line (beta version) for advanced annotators, we will be able to clarify and refine these recommendations that will later serve for less advanced annotators. We are confident that with the adoption of annotation and rendering tools, practices will evolve and stabilize so as to support new standard ways of audio-enriching videos with earcons. As a participant stated, "*Sure, it will be necessary to normalize, and then perhaps it will be possible to increase the number of earcons*". All these results show the importance of taking into account aspects related to user experiencing. They also emphasize the core issue of personalization research for accessibility.

7. CONCLUSION

In this paper we have presented the ACAV project for enhancing video accessibility on the web, together with concepts related to annotation-based enriched videos: annotations, annotations schemas and presentation models. Focusing on audio-enrichment of videos for blind people, we presented a series of questions we wanted to tackle, and two studies in which we tested different ways of combining speech synthesis and earcons. Main results show that earcons are readily perceived; that earcons and speech synthesis can be used to enhance the understanding of videos; that earcons should be accompanied with synthesized speech, prologue lexicon and explanations during the play; and that a potential side effect of earcons is related to video rhythm perception. This exploratory work is a first investigation of model-based video enrichment for accessibility, focused on the use of earcons as a way to complement speech synthesis for conveying visual information. Another key modality in annotation-based video enrichment, to be studied in the future, will be Braille display.

8. ACKNOWLEDGMENTS

This paper was supported by the French Ministry of Industry under contract 09.2.93.0966, ACAV "Collaborative Annotation for Video Accessibility" project. We thank Olivier Aubert, the main engineer of the Advene project and our mentor for their help. 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; to A. Carlier who helped us to validate the questionnaire; and to the students who provided help for data collection.

9. REFERENCES

[1] Aubert, O., Champin, P.-A., Prié, Y. and Richard, B. 2008. Canonical Processes in Active Reading and Hypervideo Production. *Multimedia Systems*, 14(6), 427–433.

[2] Baker, M. 1998. Routledge Encyclopedia of Translation Studies. Routledge. London.

[3] British Journal of Visual Impairment. 1985. The Play's the Thing - Audio description in the theatre, 3(3).

[4] Blattner, M. M., Sumikawa, D. A. and Greenberg, R. M. 1989. Earcons and Icons: Their Structure and Common Design Principles. *Human Computer Interaction*, 4(1), 11–44.

[5] Block Jr F.E., Nuutinen L. and Ballast B. 1999. Optimization of Alarms: A Study on Alarm Limits, Alarm Sounds, and False Alarms, Intended to Reduce Annoyance. *Journal of Clinical Monitoring and Computing*, 15(2), 75–83.

[6] Bonebright, T.L. and Nees, M.A. 2007. Memory for Auditory Icons and Earcons with Localization Cues. In *Proc. ICAD 2007*, 419–422.

[7] Bussemakers, M.P. and De Haan, A. 2000. When it Sounds Like a Duck and it Looks Like a Dog ... Auditory icons vs. earcons in multimedia environments. In *Proc. ICAD 2000*, 184–189.

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

[9] Choi, S. H. and Walker, B. N. 2010. Digitizer Auditory Graph: Making graphs accessible to the visually impaired. In *Proc. CHI 2010*, 3445–3450.

[10] Cohen, J. Monitoring Background Activities. 1994. In G. Kramer (Ed.), Auditory Display: Sonification, Audification and Auditory interfaces, 499–522.

[11] Crommentuijn, K. 2006. Designing Auditory Displays to Facilitate Object Localization in Virtual Haptic 3D Environments. In *Proc. ASSETS 2006*, 255–256.

[12] Edworthy, J. and Hards, R. 1999. Learning Auditory Warnings: The Effects of Sound Type, Verbal Labelling and Imagery on the Identification of Alarm Sounds. *International Journal of Industrial Ergonomics*, 24(6), 603–618.

[13] Gagnon, L., Foucher, S., Heritier, M., Lalonde, M., Byrns, D., Chapdelaine, C., Turner, J., Mathieu, S., Laurendeau, D., Nguyen, N.-T. and Ouellet, D. 2009. Towards computer-vision software tools to increase production and accessibility of video description for people with vision loss. *Univers. Access Inf. Soc.*, 8(3), 199 –218.

[14] Garzonis, S., Jones, S., Jay, T. and O'Neill, E. 2009. Auditory icon and earcon mobile service notifications: intuitiveness, learnability, memorability and preference. In *Proc. CHI 2009*, 1513–1522.

[15] Gaver, W. W. 1986. Auditory icons: using sound in computer interfaces. *Human-Computer Interaction*, 2(2), 167-177.

[16] Gaver, W. W. 1989. The SonicFinder: an interface that uses auditory icons. *Human-Computer Interaction*, 4(1), 67-94.

[17] Graham, R. 1999. Use of Auditory Icons as Emergency Warnings: Evaluation within a Vehicle Collision Avoidance Application. *Ergonomics*, 42(9), 1233-1248.

[18] Hoggan, E., Raisamo R. and Brewster, S. Mapping information to audio and tactile icons. 2009. In *Proc. of the 2009 international conference on Multimodal interfaces (ICMI-MLMI '09)*, 327–334.

[19] ITC UK. 2000. ITC Guidance on standards for audio description. Technical Report: Independent Television Commission. 16 pp.

[20] Jaskanen, S. 1999. On the inside track to Loserville, USA: strategies used in translating humour in two Finnish versions of "Reality Bites", MA essay, Univ. of Helsinki.

[21] Leung, Y. K., Smith, S., Parker, S. and Martin, R. 1997. Learning and Retention of Auditory Warnings. In *Proc. ICAD 1997*, 288-299.

[22] Mathieu, S. and Turner, J. 2007. Audio description for indexing films. *World Library and Information Congress (IFLA)*, Durban (South Africa).

[23] Mathieu, S. and Turner, J. 2007. (FRENCH) Audiovision ou comment faire voir l'information par les personnes aveugles et malvoyantes : lignes directrices pour la description d'images en mouvement. *Congrès annuel de l'ACSI*, McGill University, Montréal.

[24] Morisset, L. and Gonant, F. 2008. (FRENCH) Charte de l'audiodescription. Technical Report: Ministère des Solidarités et de la Cohésion sociale. 7 pp

[25] Murphy, E., Bates, E. and Fitzpatrick, D. 2010. Designing auditory cues to enhance spoken mathematics for visually impaired users. In *Proc. ASSETS 2010*, 75–82.

[26] Pfeiffer, S. and Parker, C. 2009. Accessibility for the HTML5 <video> element. In *Proc. W4A 2009*, 98–100.

[27] Piety, P. J. 2003. Audio description, a visual assistive discourse: an investigation into language used to provide the visually disabled access to information in electronic texts. Master of Arts thesis. Washington: Georgetown University.

[28] Piper, M. 1988. Audio Description: Pioneers Progress, *British Journal of Visual Impairment*, 6(2).

[29] Sanchez, J. and Tadres, A. 2010. Audio and haptic based virtual environments for orientation and mobility in people who are blind. In *Proc. ASSETS 2010*, 237-238.

[30] The American Council of the Blind. The Audio Description Project. Online: http://www.acb.org/adp/. Accessed 04/21/2011.

[31] United Nations. 2006. Convention on the rights of persons with disabilities. Convention: UN. New York.37 pp.

[32] W3C. Media Accessibility User Requirements. Online: http://www.w3.org/WAI/PF/HTML/wiki/Media_Accessibility_Requirements. Accessed 04/21/11

[33] W3C. Web Accessibility Initiative (WAI): strategies, guidelines, resources to make the Web accessible to people with disabilities. Online: http://www.w3.org/WAI/ Accessed 04/21/11.

[34] W3C, Web Content Accessibility Guidelines (WCAG) 2.0. Online : http://www.w3.org/TR/WCAG20/ Accessed 04/21/11.

[35] Yalla, P. and Walker , B.N. 2008. Advanced auditory menus: design and evaluation of auditory scroll bars. In *Proc. ASSETS 2008*, 105-112.