AUDIO-VISUAL RENDERINGS FOR MULTIMEDIA NAVIGATION

Tifanie Bouchara, Brian F.G. Katz,
Christian Jacquemin

LIMSI-CNRS,
BP 133, 91403 Orsay Cedex, France
tifanie.bouchara@limsi.fr, brian.katz@limsi.fr
christian.jacquemin@limsi.fr

Catherine Guastavino

CIRMMT & McGill University,
555 Sherbrooke St. West,
Montreal, QC, H3A 1E3, Canada,
catherine.guastavino@mcgill.ca

ABSTRACT

Our study focuses on multimodal information access to audio-visual databases, and evaluates the effect of combining the visual modality with audio information. To do so, we have developed two new exploration tools, which extend two information visualization techniques, namely Fisheye Lens (FL) and Pan&Zoom (PZ), to the auditory modality. The FL technique combined coherent distortion of graphics, sound space and volume. The PZ technique was designed without visual distortion but with low audio volume distortion. Both techniques were evaluated perceptually using a target finding task with both visual-only and audio-visual renderings. We did not find significant differences between audio-visual and visual-only conditions in terms of completion times. However we did find significant differences in participant's qualitative evaluations of difficulty and efficiency. In addition, 63% of participants preferred the multimodal interface. For FL, the majority of participants judged the visual-only rendering as less efficient and appreciated the benefit of the audio rendering. But for PZ, they were satisfied with the visual-only rendering and evaluated the audio rendering as distracting. We conclude with future design specifications.

1. INTRODUCTION

With the current development of computer technology, the size of data collection is rapidly increasing. Efficient methods are required to help retrieve a particular document and browse the entire collection. Research on audio-visual information access has traditionally focused on classification and indexing techniques, mainly based on content [1]. This is true for different media document, particularly image [2], audio and music [3], and video [4][5]. However once the data collection is filtered by retrieval methods, there is still a need to find efficient presentation strategies to display the query results and help browse the new dataset. The role of the interface and presentation techniques has received little attention. Our work focuses on user-centered exploration strategies to facilitate interactive information access in these datasets. Most currently available navigation methods are based on vision even for audio or audio-visual data. Our study integrates audio in browsing tools to explore multimedia collections and to determine in what extent audio modality improves exploration.

For the moment, existing systems of video browsing such as video-on-demand systems (e.g. YouTube [6] or GoogleVideo [7]), typically present the user with a simultaneous set of static fixed frame images (called key-frame or poster frame) associated with each video. As such, in these systems, the initial search effort is based on visual feedback, with the user missing the audio content. Although certain systems enable the view of the entire sequence, it is most of the time for the browsing within a unique document [8]. Only few systems can play several videos simultaneously (like the wall of *Blinkx* [9]) and they still not offer an overview of the audio content. In order to address this issue, the user should access simultaneously the audio and video content of the data.

Some auditory displays take advantage of human abilities of simultaneous listening and browsing auditory document. These systems are based on the ability to segregate sound sources played in different location (known as *cocktail party effect* [10]). The *Dynamic Soundscape* project [11] applies this concept and sound spatialization to browse a single audio file. It relies on mapping temporal position within an auditory document to spatial location so the user can listen to different portions of the audio file at the same time.

As presented in the application of Stewart et al. [12] and in the *Audio Hallway* of Schmandt [13], some other interfaces give the user the possibility to explore a collection of several sounds distributed in space around her/him without any visual feedback. On the contrary the *SonicBrowser* [14], improved in the *Audio Information Browser* [15] and the *SoundTorch* [16] are enhanced by a visual icon representation of the sounds. The user can thus browse several sound files simultaneously by navigating through a 2D soundscape. These systems exploit a concept called *aura* for *SonicBrowser* (named *torch* in [16]) consisting of circles defining the limits of user's domain of perception. All sonic objects on the perimeter or beyond are silent, while all the objects inside the disk are simultaneously played with a relative loudness depending on the distance from the center.

The concept of *aura* is derived from visualization techniques ([17], [18]) used in Zoomable User Interfaces (ZUI) (also called multiscale interfaces [19]), and particularly from the *Fisheye Lens* (FL) concept. ZUI provide a powerful way to represent and manipulate large sets of data by managing the level of detail and separating the user point of interest area (focus) from the global view (context). Among these techniques *Pan&Zoom* (PZ) relies on translations and zoom level modifications through which a homogeneous but partial view of the dataset is presented. As a focus-plus-context method, FL presents the whole dataset at a low level of detail and utilizes a movable non-homogeneous distortion (magnification) to a section of the dataset in order to examine the subset at the required level of detail. Such interfaces have been proven beneficial for visual and auditory data browsing. We proposed, developed and evaluated two novel

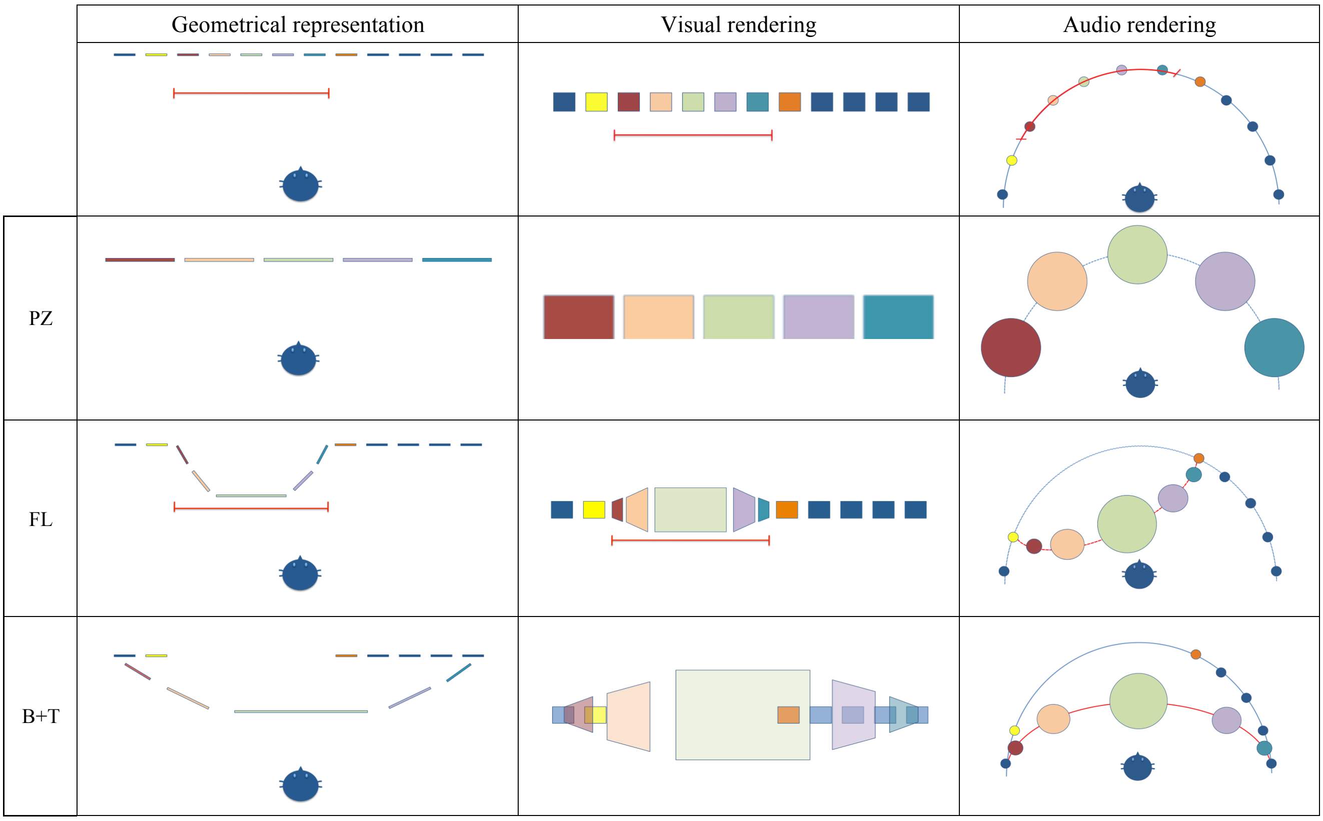

Figure 1: Schema of 3 different rendering techniques: Pan&Zoom (PZ), Fisheye Lens (FL), and Bifocal + Transparency (B+T).

audio-visual exploration techniques, combining two existing visual information access and visualization techniques, namely PZ and FL, with their auditory analogs.

The next section of this paper introduces the design of such audio-visual exploration techniques. Section 3 presents a user experiment comparing two modalities: a unimodal one (only visual) with a bimodal one (audio-visual) for the two different user interfaces, PZ and FL, while Section 4 discusses the results.

2. DEVELOPPING AUDIO-VISUAL RENDERINGS FOR NAVIGATION

2.1. Taxonomy of Zoomable User Interfaces techniques for visual information access

In Zoomable User Interfaces (ZUI) users can focus on a subset of a dataset by specifying the level of detail [17]. One of the most employed techniques is the *Pan & Zoom* (PZ). Zooming allows the user to change the scale of a specific area called *focus,* while information outside this area is discarded. Panning allows the user to translate the viewport. In such an approach, the rendering is homogeneous (without distortion) but there is no global view. As users cannot see the relationship between the visible portion and the entire structure, they can be disoriented by the lack of visual *context*. On the contrary *Focus-plus-context* techniques, combine the focus area and the global view in a single display. Among these techniques *Bifocal Display* superimposes the focus area over the context. Both areas are presented without distortion but the focus masks a part of the context and some information cannot be displayed.

Another option is to distort the rendering as in the *Fisheye Views* [20] (see also [18] for a review on distortion-oriented techniques). Originally this technique consisted of the suppression of non-interesting part of the information according to a threshold. It relied on the calculation of the *Degree Of Interest* (DOI) for each object and was designed for hierarchical information. An improvement of this method was designed for tree structures with the concept of *Hyperbolic Browser* [21] where more space is assigned to a portion of the hierarchy while still embedding it in a much larger context. The concept was also extended to a graphical fisheye lens in [22]. The focus area is enlarged while the rest of the image is reduced proportionally to the Euclidean distance to the center of the lens. This method combines the accuracy of spatial distortion while preserving the simultaneous visualization of the focus and context areas.

Pook et al. [23] suggested a transparency method where the contextual view is a transparent layer drawn over the magnified focus of attention. There is no masking and no distortion, however the large amount of information presented simultaneously results in more efforts for the user to distinguish one view from another.

2.2. Extension to audio-visual renderings

To extend the visualization techniques to the auditory domain and design audio-visual browsing methods, we chose to map different properties of graphical rendering to audio rendering: position of the objects are map from visual position in the screen picture to the spatialized audio rendering and size of the objects are mapped to the sound level. The mapping is presented in Fig.1. for 3 different techniques: Pan&Zoom (PZ), Fisheye Lens (FL), and Bifocal+Transparency (B+T).

The link between graphical and audio space can be seen as projection from the geometrical Cartesian representation (Fig.1 col.1), corresponding to a top-view of the visual rendering (Fig.1 col.2), to a polar representation for audio rendering (Fig.1 col.3). Indeed it is equivalent to say that the graphical rendering is analogous to the front space of spatialized audio rendering. The objects' positions are also coherent, but not congruent, between graphical and auditory renderings.

The mapping between the visual size and the volume of the object's sound is inspired from real life as both are linked to the distance from the user. Thus, we considered that the larger an object is in the visual rendering, the louder the sound of this object must be.

In the first method extending Pan&Zoom, there is no distortion. Also the objects have a homogeneous size and volume and are uniformly distributed in space. The main problems are that only few objects are displayed and no context can be perceived.

The FL design uses a visual position distortion corresponding to an angle manipulation for the spatialization of sounds. The progressive graphical magnification is equivalent to a progressive audio level increase. The main advantage for both modalities is the presence of context. However this results in a graphical distortion that can disturb users and in an audio distortion that can be difficult to perceive, because of the small azimuthal distortion. Indeed it is quite difficult to segregate the different sources inside the lens, as the objects are close to one another.

In the third method B+T, we decided to improve segregation of sound sources inside the focus area. Sound sources should be more spread out so that users can better segregate multiple audio sources [24]. B+T also combines bifocal display and transparency method. The rendering is also similar to FL but the focus area is centered and transparently superimposed on the context. The sources are more distinguishable, however some sources are heard as located in the same direction because of superposition.

Finally graphical and audio renderings can be combined non congruently, associating the graphical rendering from one method with the audio rendering from another, e.g. a PZ visual rendering with a B+T audio rendering.

2.3. Visual rendering implementation

Two graphical renderings methods were implemented: Fisheye Lens (FL) and Pan&Zoom (PZ). They are processed through shaders, small programs that are run on the graphics card [25].

The PZ technique renders only a single portion of the environment. This method corresponds to the manipulation of a camera as described in space-scale diagram by Furnas and Bederson [19]. The camera can be moved through the left/right axe (panning), and through the back/forward axe to change the scale of detail (zooming). Thus only a part of the global view is captured then enlarged to obtain an image of desired size, i.e equal to the screen size.

The FL rendering is divided into three parts: in the center area of the lens, objects are homogeneously magnified, outside the lens objects' size is not modified while in between the size is progressively interpolated. To compute the rendering of FL, three passes are necessary.

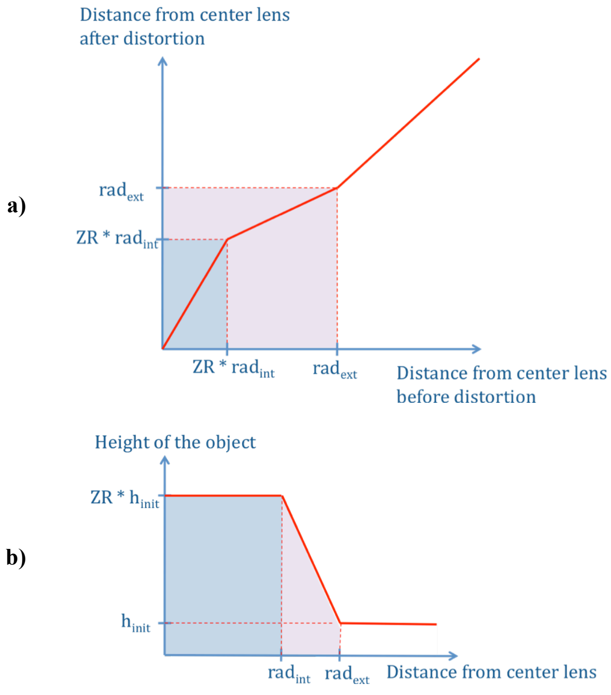

Figure 2: Distortion curves for FL technique. a) Position of objects after distortion. b) Height of object according to their visible position.

```
  if p ∈ d1
     then textFinal <- textZoomed
     else if p ∈ d2
        then textFinal <- f(textZoomed,ZR,p)
        else if p ∈ d3
           then textFinal <- g(textNorm,ZR,p)
           else textFinal<- textNorm
        end if
     end if
end if
```

Figure 3: Pseudo-code of the lens shader.

Then the rendering is processed by taking a zoomed view of the environment and stored as a second texture (*textZoom)* enlarged so that its size equals the screen size. The environment part captured in zoomed view is the one contained inside the lens of radius *rad$_{ext}$*. Both textures are then mixed and distorted according to the fisheye strategy: in the focus area (of radius *rad$_{int}$*) the zoomed view is used to have magnification without pixellization (when screen size is lower or equal to the texture size), while the normal view is used for the context part. The shader is parameterized to select the parts of the texture that are enlarged and to define the strength of the distortion according to a distortion ratio *ZR*. The deformed texture *textFinal* is then mapped to a quad parallel to the projection plane and is finally rendered on a view port that covers the whole display screen. A white border marks the boundary of the lens and allows the user to easily locate the position of the lens even at low distortion levels. Figure 2 presents the distortion curves chosen to modify the position or the size of the objects in the graphical rendering of FL. Figure 3 presents the pseudo-code for the lens shader creating the distortion: for each pixel *p* of the rendering picture we allocate the corresponding texture depending on which zone *p* belongs to (center, outside or between). *d1, d2* and *d3* are discs delimiting each part of the lens. The radius of the discs are rad_{int}, $(rad_{int+}rad_{ext})/2$, and rad_{ext} respectively. *f* et *g* are two functions of distortion using the curve b (Figure 2).

2.4. Audio rendering implementation

For audio rendering, we considered that a spatial separation among sound sources is necessary for perceptual segregation. The implemented audio rendering was also the bifocal+transparency (Fig. 1). We adapted our audio B+T technique to work with PZ or FL graphics described in Fig 1. The multimodal congruency is thus not respected but the link between audio and graphical renderings is still coherent.

As there is no visual distortion with PZ, we tried to keep a homogeneous rendering for audio. There is no azimuthal distortion in this audio rendering as illustrated in (2). However, we applied a low distortion on the volume (*vol*) to reduce the number of sources played simultaneously (1). The volume distortion is similar to the FL volume distortion (3) but with a maximal lens radius, i.e. equal to the width of the window rendering (screen size if in fullscreen mode).

$$vol = \begin{cases} v_{\min} + \log(ZR), & if \quad |dz| < rad_{\text{int}} \\ v_{\min} + \log(ZR) * e^{-c*|dz - rad_{\text{int}}|}, & if \quad rad_{\text{int}} < |dz| < rad_{ext} \\ v_{\min}, & if \quad rad_{ext} < |dz| \end{cases} \quad (1)$$

$$az_{bis} = az \quad (2)$$

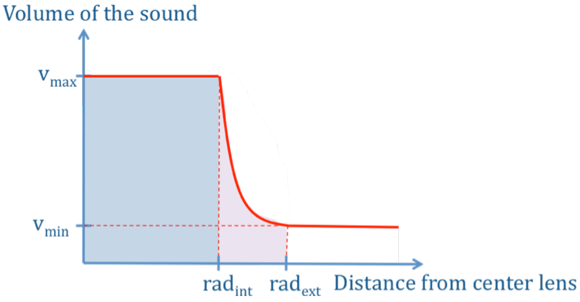

Figure 4: Nonlinear distortion curve used for volume in audio renderings from the visual position of objects.

For FL, the process relies on a distortion on both position and size in graphics and also on both angular position (*az*) and volume (*vol*) in audio. Equations of the proposed auditory distortion are presented in (3) and (4). *dz* represents the visible distance, after visual re-processing, between the object and the center of the lens. *ZR* is the zoom ratio or magnifying scale. v_{min} is the minimal level when no magnification is applied or when sources are out of the focus area. *c* is a constant giving the attenuation of volume between focus and context areas. α_{int} and α_{max} represent azimuth of sources on the internal perimeter rad_{int} and on the external perimeter rad_{ext}. Figure 4 presents the distortion curves chosen to modify the volume of objects according to their graphical position.

$$vol = \begin{cases} v_{\min} + \log(ZR), & if \quad |dz| < rad_{\text{int}} \\ v_{\min} + \log(ZR) * e^{-c*|dz - rad_{\text{int}}|}, & if \quad rad_{\text{int}} < |dz| < rad_{ext} \\ v_{\min}, & if \quad rad_{ext} < |dz| \end{cases} \quad (3)$$

$$az_{bis} = \begin{cases} dz * \dfrac{\alpha_{\text{int}}}{rad_{\text{int}}}, & if\, |dz| < rad_{\text{int}} \\ A * dz + B, & if \quad rad_{\text{int}} < |dz| < rad_{ext} \\ with \quad A = \dfrac{\alpha_{\max} - \alpha_{\text{int}}}{rad_{\max} - rad_{\text{int}}} & and \quad B = sg(dz) \\ az, & if \quad rad_{ext} < |dz| \end{cases} \quad (4)$$

2.5. Sound spatialization technique

Sounds are spatialized through a *virtual Ambisonics* technique for the auditory part of our bimodal interface [26]. This mixed method between Ambisonic encoding and binaural decoding allows us to treat simultaneously a large amount of sources without latency while providing a rendering on headphones usable for general public.

However, as the decoding part is independent from the encoding, the diffusion system could be replaced by a more immersive system with loudspeakers like VBAP, Ambisonic or WFS. Furthermore, we chose to use 2D audio renderings in this study but the implementation offered the possibility to extend the methods to 3D sound spatialization.

2.6. Global architecture

The software architecture (Fig. 5) was based on the *SceneModeler* package designed in two different parts: a virtual scene descriptor and a sound spatializer [27]. We used a triangular structure where all vertices (user, visual and sonic components) are connected by interaction links. The scene descriptor tool and the spatializer communicated through OSC messages via UDP protocol [28].

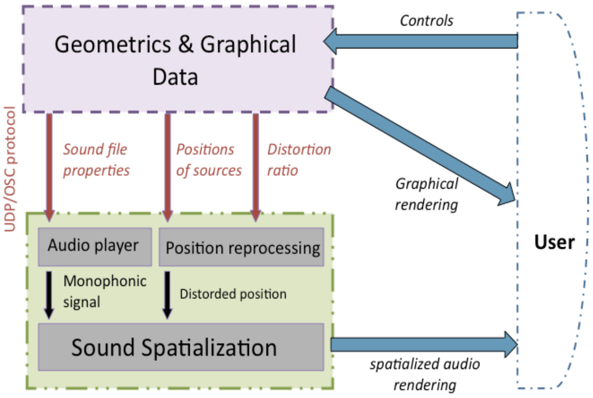

Figure 5: Structure of the interface.

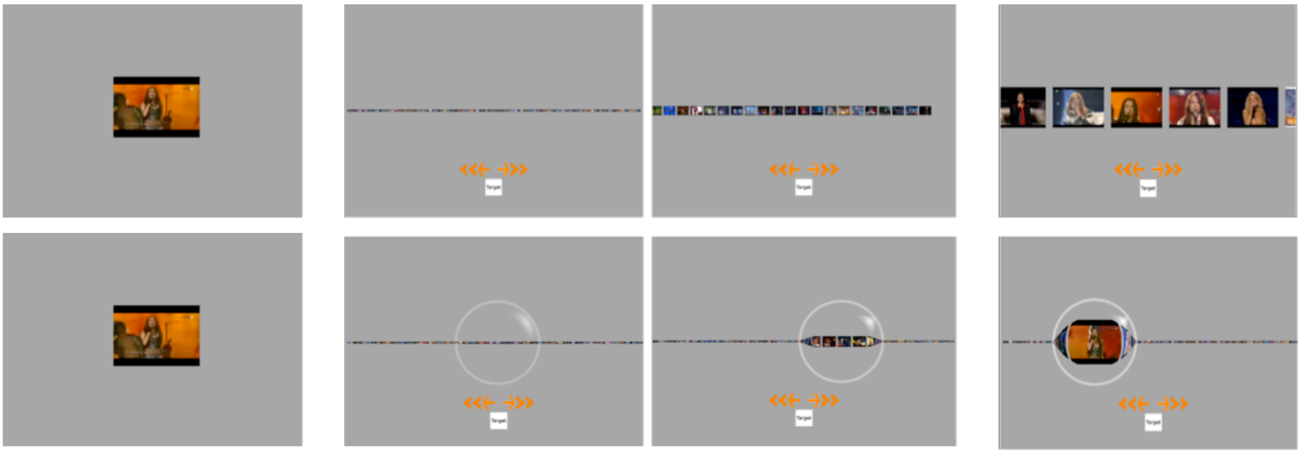

Figure 6. Storyboard of the task with both methods: PZ (top) and FL (bottom).

3. EVALUATION

The aim of the study is to evaluate the possible contribution of audio to browse audio-visual databases. Thus, the experiment is based on a comparison between audio-visual (AV) against visual-only (V) renderings. Two different navigation techniques, Pan&Zoom (PZ) and Fisheye Lens (FL), were tested in order to assess whether the audio influence could be depending in the chosen visual or audio rendering technique.

3.1. Experimental protocol

Participants

Sixteen participants, with basic computer skills and familiar with the use of a mouse, took part in this experiment (11 males, 5 females, mean age 27). They received $15 for their participation.

Design and conditions

We used a 2x2 within-subjects factorial design with 2 modal conditions (V/AV) * 2 methods (PZ/FL).

The graphical rendering of the FL was based, as described earlier, on a Fisheye Lens distortion. The radius of the lens was rad_{ext}= 178 px for a screen size of 1270x940 pixels. The visual rendering was also divided into three parts: the center of the lens (rad_{int} = 2/3* rad_{ext}) was magnified in a heterogeneous way, the external area was a global view and a progressive distortion was used in between the two. For audio-visual presentation, the audio rendering corresponded to the audio bifocal transparency described in (3) and (4). Outside the lens, sources were spatially spread out between -α_{max} and α_{max} = 90°. Sources inside the lens were spread between -α_{max} and α_{max} = 90° too but with the sources belonging to the internal area of the lens spread between -α_{int} and α_{int}=70°. With the highest deformation ratio (ZR=20), three different sources sufficiently separated in space could be heard simultaneously.

No distortion was used for the visual rendering of PZ. The audio rendering was based on the same audio distortion as for FL with a lens radius equal to the width of the screen, i.e. rad_{ext}=1270 px. More sound sources could be heard simultaneously than for FL, up to six or seven sources with the highest zoom ratio.

Hypothesis

Audio rendering can convey redundant information to reinforce visual feedback, or convey additional information to complement visual information. Therefore, we hypothesized that the addition of redundant and complementary audio rendering would enhance navigation and information when browsing an unorganized audio-visual collection. In addition, we investigate the effect of the rendering technique itself, and hypothesize that a more focused audio rendering (used in FL), i.e. with few but relevant sound sources, would be more useful that a less focused audio rendering (used in PZ).

The video collection

To evaluate renderings in a realistic context, we used a collection of 100 video clips from the Eurovision Song Contest from 2005 to 2008 [29]. The videos were selected from the result set of the textual query “Eurovision” on the video-on-demand system YouTube [6]. The video clips were excerpts of singers' performances, each showing a different singer and a different song. For each video, we extracted a 10-second clip corresponding to a musical phrase. Video clips were then played in a loop. The different clips could easily be distinguished through the visual properties of the singers, their voice and the musical genres provided several clues for identification. Moreover there was a good balance between visual and auditory cues for identification and a consistency between simultaneous visual and auditory components. Finally the videos were selected from the same TV program to ensure homogeneity of the collection.

Videos were stored with a 160x120px size and displayed at 11x8px before magnification. The soundtrack of the videos were extracted from the movie and stored as monophonic signal (left channel only) in 44,1kHz in 16 bits wav files. They included singing voice and instrumental music. To spatialize the sounds we considered that each sound file was attached to the center of the corresponding visual object. All stimuli and audio-visual rendering examples are available on the web [30].

Retrieval task

The task was to watch a video clip and then browse the video collection to retrieve it as quickly as possible. Each trial was divided into three steps represented in Figure 6 : a presentation of the targeted movie, then a step of exploration to find the target by changing scale or distortion level and position of the focus, and finally the selection of a movie with the user clicking on it.

Participants started by clicking on a button to see and listen to the target, a 10 second video clip presented in isolation once with no distortion. At the beginning of the exploration step, the user was presented with an overview of the 100 videos in the collection, arranged in a line in random order at a reduced size and sound level, so that the user could not discern the different clips in this view. The user had to use the zoomable techniques proposed. The minimal size of the videos on the 1270x940 screen is 11x8 pixels while the maximal size is 220x170 pixels. As the videos are very small, thousands of stimuli would have been necessary to fill in the screen resulting in hours of browsing experimental sessions. Hence the line arrangement was preferred.

Procedure

After a training block (on all 4 conditions), the actual experiment was divided into four blocks corresponding to the 4 conditions of the factorial design, namely AV-FL; V-FL; AV-PZ; V-PZ, presented in counterbalanced order using a Latin square design. Each block consisted of 15 trials. On each trial the presentation of the videos was randomized and a new video clip was randomly chosen as a target.

After each block, participants were asked to provide free-format comments and to evaluate for each condition: the perceived efficiency, adaptability, and difficulty. After the experiment, participants were asked to indicate their preferred method, audio-visual condition and combination.

The entire experiment lasted around one hour and half per participant. Participants were invited to take breaks after each block.

Apparatus

For faster computing, we used a distributed multi-platform architecture on two different computers for this experiment. The first one processed only the audio rendering while the second one managed with navigation and graphical rendering. We used the platform VirtualChoreographer on a AMD Athlon 64X DUAL CORE 5000+ 2.60 GHz with a Nvidia 8600T graphic card for the navigation process and graphical rendering and the Max/MSP environment on a MacBookPro 2.4Ghz with an integrated digital sound card for the audio display. The audio rendering was presented on AKG K271 headphones.

3.2. Results

Our dependant variables included completion times, number of errors, adaptability, difficulty and efficiency ratings, collected after each block, as well as overall preference ratings as free format descriptors collected at the end of the experiment. To present the different results we used a color code throughout the paper: PZ conditions are represented in green, FL in blue, and audio-visual conditions are shaded in.

For the statistical analysis we first removed miss trials for which a wrong video was selected (~3.2% among the 960 trials: 5 errors for AV-PZ, 2 for V-PZ, 14 for AV-FL and 10 for V-FL; 240 trials for each condition). Then we removed outliers from the hit trials for each condition and participant (13 outliers for AV-PZ, 11 for V-PZ, 6 for AV-FL and 7 for V-FL). Outliers corresponded to hit trial for which the completion time was more that two standard deviations away from the mean. Completion times were considered only for hit trials.

A 2*2 factorial ANOVA revealed that completion times were significantly lower for PZ than for FL (F(1,890)=8.82; p=0.003) (see Fig. 7). No interaction effect between methods and modality were observed (F(1,888)=0.06; p=0.81). We subsequently report the comparison between AV and V conditions for each method separately.

For the PZ method, no significant effect of modality on completion times was observed (F(1,448)=0.46, p=0.59). However, the analysis of subjective ratings (Fig. 8 and 9) and the free-format comments indicated that participants evaluated the addition of audio rendering negatively for PZ technique. Indeed they rated V-PZ as significantly more easier to use than AV-PZ (t(15)=2.07, p=0.05). V-PZ was also perceived as more efficient than AV-PZ but this difference did not reach statistical significance. In addition, participants commented that PZ method “produced too much overlapping noise when scanning many videos” which is “more a distraction than an aid”. They further commented on the difficulty to associate the sound to the right movie as too many sounds were presented at once. Thus, even in the audio-visual condition, the PZ method was “mostly a visual scan instead of audio-visual” all the truer, as visual information is highly reliable in this technique without distortion. Participants also enjoyed the visual rendering providing “visual scanning of many items of the same size” and “like the uniformity when scrolling”.

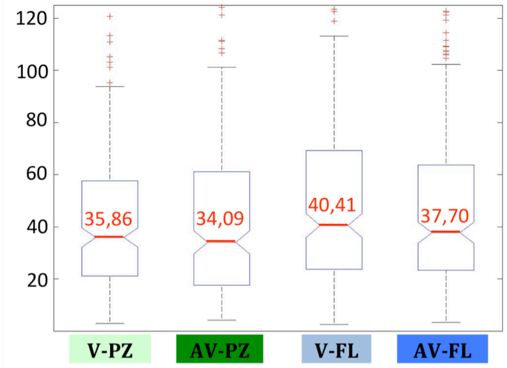

Figure 7. Mean completion times in sec. collapsed over all tasks and participants and grouped by conditions.

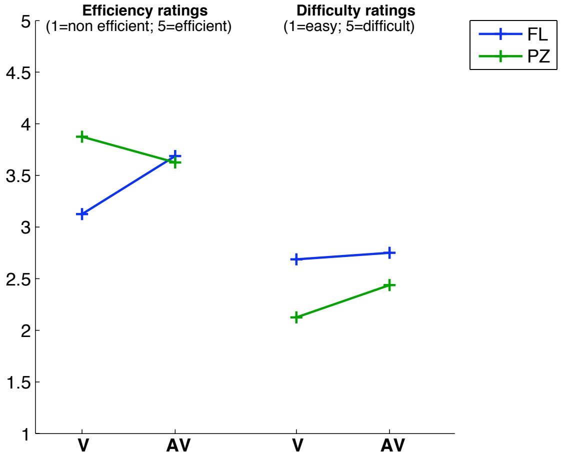

Figure 8. Average of subjective ratings collapsed over all tasks and participants and grouped by conditions.

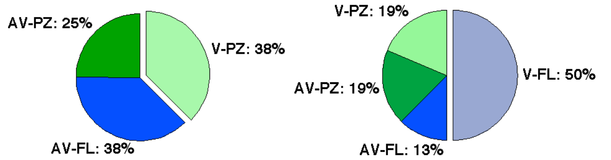

Figure 9: Participants' preferences (N=16): left) most enjoyed combination, right) less enjoyed combination.

As for PZ, while completion times for AV-FL modality were lower than for V-FL (37.70 sec. 40.41), the difference did not reach statistical significance (F(1,442)=0.06,p= 0.80). However the qualitative results (Figures 8 and 9) and comments differ from those of PZ as modality does not affect difficulty while audio improves the perceived efficiency for FL (t(15)=2.52, p=0.02). Participants justified that FL displayed too many sounds, but less than PZ, and that sound was also beneficial when zoomed in: "At first, the audio seemed to be a distracter, but once the magnifier is zoomed in, it is helpful". Thus the lens allowed the user to "visually scan multiple videos while audio scanning just few". In this case audio is used to compensate for the insufficiency of visual information as the "view of zoomed-in-image is limited" with FL.

To summarize, in terms of techniques, PZ is faster than FL and in terms of modality, the addition of audio rendering has a positive effect for FL and a negative effect for PZ. In one of the participants' own words "With PZ there are too much noise but with FL it's funny you've got only 3 or 4 sounds. But it is easier with PZ cause you can see all the videos". In their free comments, 81% of the 16 participants reported relying on audio during the experiment, either to browse sonically the video collection (55%) or only to confirm the visual selection certain ambiguous videos (25%). displays the conditions preferred and most disliked by participants. The majority of participants (63%) preferred bimodal conditions. Similarly, a similar percentage (69%) of participants disliked unimodal conditions. Together, these findings indicate the addition of audio rendering enhances user experience.

4. CONCLUSION

This study aimed to evaluate if the addition audio rendering could improve navigation in audio-visual collections using a multimodal user interface. The first step of the study was to suggest ways of combining audio rendering with existing graphical rendering. Two audio-visual methods related to Pan&Zoom and Fisheye Lens have been implemented in a visual-only mode and an audio-visual mode. They were evaluated with respect to the contribution of audio on video browsing. No significant differences were observed between multimodal and purely visual interfaces in terms of completion times. This could be explained by the predominance of vision in human perception but also by participants' previous experience with visual searching while audio rendering is rarely used for navigation. However, subjects self-reported audio as an enjoyable and interesting way to provide additional information. So we believe that the absence of performance improvement due to the inclusion of audio could be due to compensation between the positive effects (redundant and complementary information transfer) and some negative effects (auditory fatigue and discomfort). Participants also reported the background noise produced by the contextual sources as "annoying". In future instances, to avoid auditory fatigue due to the presentation of non-relevant sounds, we suggest keeping silent all sources outside of the lens (reciprocally outside the screen for PZ) as done by [14] and [16].

Furthermore participants reported a preference to rely mainly on visual rendering for navigation, and for a graphical rendering without distortion with several videos presented at the same time with homogeneous magnification. Participants' ratings and comments reveal the positive effects of conveying information through the auditory modality when focused on few sound sources as in the FL case. Our results suggest also us to design audio-visual renderings differently to benefit from advantages of both modalities. Providing a combination of the homogeneous PZ visual rendering plus the distorted FL audio rendering focusing on few sound sources should improve the navigation step.

Our primary goal was not to compare PZ to FL technique, as we focused mainly on the addition of audio renderings. However our results show that PZ significantly outperformed FL both in terms of completion times and affective reactions. Even thought the same control was used for navigation with PZ and FL, selecting a video might have been more difficult with FL as the lens had to be positioned on the video to select it. With PZ on the other hand, participants could click on and thus select any video displayed on the screen. The advantage observed for PZ could therefore possibly be attributed to interaction control.

Finally, this audio technique proposed here could be improved further by including additional spatial auditory cues to segregate sound sources, particularly elevation. For instance, we could apply azimuthal distortions in the same manner and arrange multimedia objects using a grid instead of a straight line – which is more representative of a real application. The tools could also be extended to an immersive 3D environment. However PZ is a non-egocentric concept that is not really suitable to immersive 3D scenes. On the contrary FL could be interesting to explore these environments.

Participants' positive reactions during the experiment showed the beneficial effect of audio rendering when focused on a limited number of sound sources (3 or 4 at a time). Future studies will investigate other audio design methods for multimodal navigation. Sound level distortion could be combined efficiently with distortion of other sound parameters to increase the effect. Specifically, a simulation of distance and presence, by adding reverberation or varying the high-low frequencies balance, could be used to differentiate foreground and background sound sources, thus directing attention to relevant sound objects and improving audio selection.

5. REFERENCES

[1] M. S. Lew, N. Sebe, C. Djeraba and R. Jain, "Content-Based Multimedia Information Retrieval: State of the Art and Challenges." *ACM Trans. Multimedia Comput. Commun. Appl*, vol. 2(1), pp. 1–19, 2006.

[2] A. Halawani, A. Teynor, L. Setia, G. Brunner, H. Burkhardt., "Fundamentals and Applications of Image Retrieval: An Overview." in *Datenbank-Spektrum, vol. 18*, pp. 14-23, August 2006.

[3] J. T. Foote, "An Overview of Audio Information Retrieval." in *ACM Multimedia Systems*, vol. 7 (1), pp 2-10, January 1999.

[4] R. C. Veltkamp, H. Burkhardt and H.-P. Kriegel (eds.), *State-of-the-Art in Content-Based Image and Video Retrieval*. Kluwer, 2001

[5] O. de Rooij, C.G.M. Snoek and M. Worring, "Query on Demand Video Browsing." in *Proc. of the ACM Int. Conf. on Multimedia (MM'07)*, Augsburg, Germany, 2007, pp. 811-814

[6] http://www.youtube.com

[7] http://www.video.google.com

[8] W. Hürst, "Interactive audio-visual video browsing." in *Proc of the 14th ACM Int. Conf. on Multimedia (MULTIMEDIA '06*), 2006, pp. 675-678.

[9] http://www.blinkx.com

[10] B. Arons, "A Review of the Cocktail Party Effect. *J. of the American Voice I/O Society,* 12, pp. 35-50, 1992.

[11] M. Kobayashi and C. Schmandt, "Dynamic Soundscape: mapping time to space for audio browsing" in *Proc. of the Conf. on Human Factors in Computing Systems (CHI '97), New York, NY, 1997, pp. 194-201.*

[12] R. Stewart, M. Levy and M. Sandler, "3D Interactive Environment for Music Collection Navigation" in *Proc. of the 11th Conf. on Digital Audio Effects* (DAFx-08), Espoo, Finland, 2008. pp. 13-17

[13] C. Schmandt, "Audio Hallway: a Virtual Acoustic Environment for Browsing" in *Proc. of the Symp. on User Interface Software and Technology (UIST'98), 1998, pp.* 163-170

[14] M. Fernström, and E. Brazil. "Sonic Browsing: an auditory tool for multimedia asset management" i*n Proc. of the 7th Int. Conf. on Auditory Display (ICAD'01)*, Espoo, Finland, 2001, pp. 132-135.

[15] E. Brazil, M. Fernstroem, G. Tzanetakis, and P. Cook, "Enhancing sonic browsing using audio information retrieval" in *Proc. of the 8th Int. Conf. on Auditory Display (ICAD2002),* Kyoto, Japan, 2002

[16] S. Heise, M. Hlatky and J. Loviscach, "Aurally and visually enhanced audio search with soundtorch" in Proc. Proc. of the 27th Int. Conf. on Human Factors in Computing Systems, 2009, pp. 3241-3246

[17] A. Cockburn, A. Karlson, and B. B. Bederson, "A Review Of Overview+Detail, Zooming, And Focus+Context Interfaces". *ACM Computing Surveys (CSUR), vol. 41 (*1), 2008.

[18] Y. K. Leung and M. D. Apperley "A review and taxonomy of distortion-oriented presentation techniques" in *ACM Transactions. on Computer-Human Interaction (TOCHI), vol. 1(2), pp. 126-160, 1994.*

[19] G. W. Furnas and B. B. Bederson, "Space-Scale Diagrams: Understanding Multiscale Interfaces" in *Proc. of the Conf. on Human Factors in Computing (CHI '95), 1995,* pp. 234-241.

[20] G. W. Furnas, "Generalized Fisheye Views" in *Proc. of the Conf. on Human Factors in Computing Systems (CHI'86), 1986,* pp. 18-23.

[21] J. Lamping, J., R. Rao and P. Pirolli, "A Focus+Context technique based on hyperbolic geometry for visualizing large hierarchies" in Proc. of C*onf. on Human Factors in Computing Systems (CHI' 95), 1995.*

[22] M. Sarkar and M. H. Brown, " Graphical Fisheye Views. *Communication of ACM*, vol. 37 (12), *pp.* 73-83, 1994.

[23] S. Pook , E. Lecolinet, G. Vaysseix and E. Barillot. Context and interaction in Zoomable User Interfaces" in *Proc. of the 5th Int. Work. Conf. on Advanced Visual Interfaces (AVI 2000)*, 2000, pp. 227-231

[24] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990.

[25] R. Fernando and M. J. Kilgard*, The Cg Tutorial: The Definitive Guide to Programmable Real-Time Graphics,* NVIDIA, 2003. Available on-line: http://developer.nvidia.com/object/cg_tutorial_home.html

[26] M. Noisternig, A. Sontacchi, T. Musil and R. Holdrich, "A 3D Ambisonic Based Binaural Sound Reproduction System." in *Proc. 24th Int. Conf. of the AES: Multichannel Audio, The New Reality,* 2003. *pp. 1-5*

[27] T. Bouchara, "Le SceneModeler: des outils pour la modélisation de contenus multimédias interactifs spatialisés" in *Proc13ème Journées d'Informatique Musicale (JIM'08)*, GMEA-AFIM, Albi, France, 2008. pp. 8-13

[28] M. Wright, A. Freed and A. Momeni, "OpenSound Control : State of the Art 2003" in *Proc. of the 2003 Conference on NIME*, Montreal, Canada, 2003. pp. 153-159

[29] http://www.eurovision.tv

[30] http://www.limsi.fr/Individu/tifanie/downloads/audiovisual_renderings.zip