# AB – Web : Active audio browser for visually impaired and blind users

Patrick Roth1, Lori Petrucci1, André Assimacopoulos2, Thierry Pun1

(1) Department of Computer Science, University of Geneva
24, rue du Général Dufour
CH – 1211 Geneva 4, Switzerland
(2) UCBA – Swiss Central Union of and for Blinds
Schützengasse 4
CH – 9000 St.-Gallen, Switzerland

## Abstract

The Internet now permits easy access to textual and pictorial material from an exponentially growing number of sources. The widespread use of graphical user interfaces, however, increasingly bars visually handicapped people from using such material. In this context, our project aims at providing sight handicapped people with alternative access modalities to pictorial documents. More precisely, our goal is to develop an augmented Internet browser to facilitate blind users access to the World Wide Web. The main distinguishing characteristics of this browser are : (1) generation of a virtual sound space into which the screen information is mapped; (2) transcription into sounds not only of text, but also of images; (3) active user interaction, both for the macro-analysis and micro-analysis of screen objects of interest; (4) use of a touch-sensitive screen to facilitate user interaction. Several prototypes have been implemented, and are being evaluated by blind users.

**KEYWORDS**: WWW, blind user access, sound space, image analysis.

# INTRODUCTION

## Context

Internet access, coupled with graphical user interfaces (GUI) and browsers, has become common in education, business, and at home. However, due to the widespread use of GUIs, the so-called enabling technologies for sighted have become disabling technologies for the visually-impaired. Blind users are less and less able to benefit from the enormous wealth of archived digital multimedia that is offered over Internet.

The development of electronic aids for visually-impaired persons has been ongoing for several decades, such as for reading, facilitating mobility, and for educational and occupational activities; see e.g. [2] for a review. More recently, a pressing need to offer blind users access to Internet-based digital information has been identified (e.g. [10] [11]). The major difficulty in designing Internet browsers for blind people stems from the essentially *bidimensional* layout and nature of the textual and pictorial information that has to be presented. This is to be contrasted with the essentially "one-dimensional" nature of existing output devices such as Braille lines or text-to-speech converters. Another difficulty in presenting digital information to blind people users arises from the presence of embedded images; they often bear essential information, and should be suitably presented to users.

## State of the art

Current browsers for blind users typically transform the bidimensional content of the WWW documents into spoken text, for example by analyzing the incoming HTML source code (e.g. [7] [8]). These browsers often have difficulties in presenting the global layout of documents, and usually remove all pictorial information. In this context, we present here the design principles of an augmented WWW browser for blind users that addresses these issues.

The original aspect of our project consists in the creation of an immersive environment (3D sound space) into which the HTML document is mapped. In this environment, each object (text, link, image) is represented by an earcon (e.g. [6] [13]) whose simulated 3D location depends on the location of the displayed element in the Web browser window. The blind user can therefore explore the spatial organisation of the document in order to obtain a mental image of what is displayed. Moreover, all images that are included in the document are also translated into sound. This translation should help blind users to discover the global characteristics of the displayed image during the exploration phases. These design principles have led to a working prototype currently being evaluated by visually handicapped people, using HTML documents of varying complexities (see Figure 1).

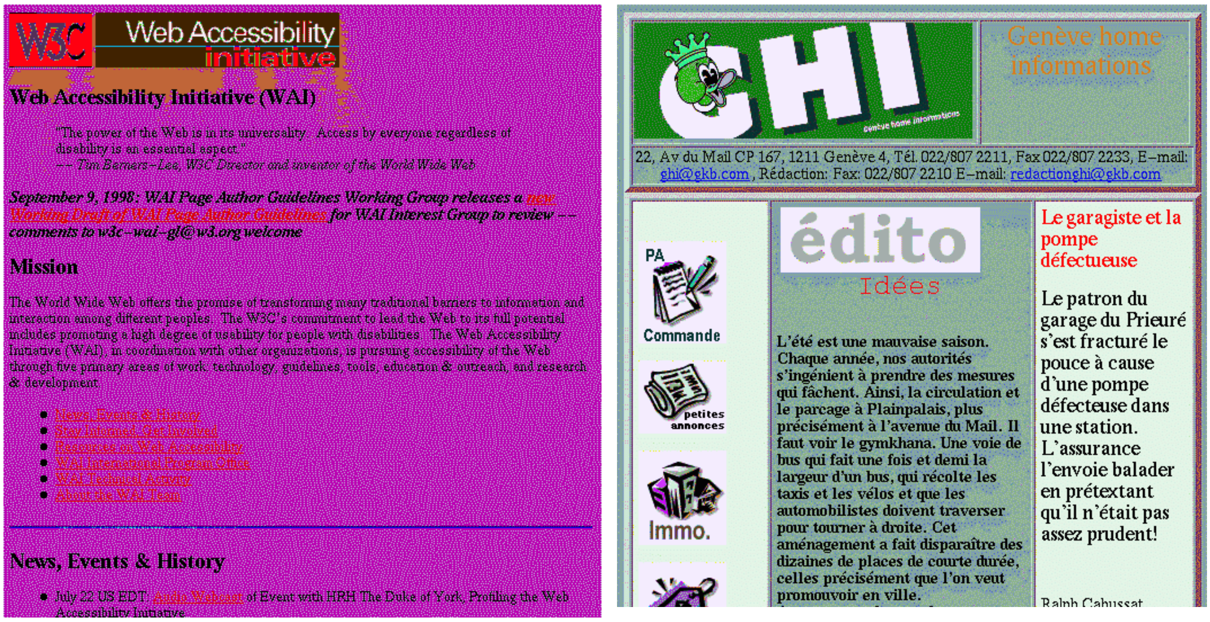


**Figure 1: Examples of HTML documents used for testing.**

# USER INTERACTION

## Macro-analysis and micro-analysis of a document

Two successive exploration phases are performed by users in order to analyze a WWW document. These phases correspond to the "where" and "what" stages in visual perception [9], and are supported by the browser. First, similarly to the "where" stage of vision perception, a *macro-analysis* phase allows the understanding of the document structure and of the element types (e.g. text, images, forms) displayed in the browser viewport. Secondly, as with the "what" stage, a *micro-analysis* phase lets users focus on one particular object to obtain its information content. These phases are iterated: users interactively perform a succession of macro- and micro-analysis explorations of the viewport. The content and position of each HTML elements can be retrieved by obtaining their CSS attributes [3] using JScript.

## Active interaction for document exploration

It is well known that interaction between a user and a system is essential for the understanding of the system's functionalities. Furthermore, the "action-perception loop" linking user actions and system responses, has to be as short as possible. To facilitate interaction, user input is accomplished via a *touch-sensitive screen*, that eliminates the need for inserting a pointing device (such as a mouse) into the loop. Using fingers, the user actively explores the screen both for macro-analysis, with the system responding with an audio description of the screen layout and elements relationships, and for the micro-analysis, with the system transcribing into sounds the information pertaining to the content of an element of interest.

## Audio rendering by means of a virtual sound space

Audio feedback in response to active user exploration is transmitted using a stereophonic headset. We have studied two approaches for this audio transcription of information. The simpler technique consist of using a text-to-speech converter, for reading text and for describing the boundaries positions of a given screen element.

To provide more elaborated feedback allowing e.g. to "display" screen layout and images, we are investigating a second approach consisting of the generation of a *virtual sound space* [4] [12]. Such a sound space is an immersive environment in which a particular acoustic signal $S(t)$ can be perceived as originating from a given spatial location (typically expressed in azimutal angle $\theta a$, elevation angle $\theta e$ and distance $d$, all defined w.r. to the listener). This approach relies on Head-Related Transfer Function (HRTF) [1]. HRTFs model the spectral modifications of the source signal $S(t)$ due to the listener's external ears, as a function of location $\{\theta a , \theta e , d\}$ of $S(t)$. HRTFs depend on each individual, and come in pairs $\{HL, \theta a , \theta e ,d\ (),HR, \theta a , \theta e ,d()\}$ corresponding to the filtering effects of the left and right ears respectively. The stereophonic effect of a sound source $S(t)$ at any given location can thus be simulated by convolving $S(t)$ with the appropriate HRTFs. For audio rendering of an element at a given location in the screen location, $S(t)$ will be related in the macro-analysis case to the type of the element, and in the micro-analysis case to its content; $\{\theta a , \theta e ,d\}$ will be related to the element's position. In this way, it is possible to generate an immersive virtual sound space related to a particular Web document. Concretely, we use for our project the Intel library RSX [5] which is based on HRTF technology.

# INFORMATION ENCODING

## Document elements to be encode

The document items requiring audio encoding are: *document structure* and *element types* during the macro-analysis phase, and *information content of elements* during the micro-analysis phase. Object locations and types (e.g. text paragraph, image) are determined during macro-analysis by parsing the HTML source code of the document. During micro-analysis, the information that requires encoding can be either textual or pictorial, where a picture is a two-dimensional array of pixels $P(x,y)$. {x,y} are coordinates expressed with respect to an arbitrary image origin $O$, and $P$ is a pixel attribute (e.g. grey-level) or set of pixel attributes (e.g. RGB or HLS values) of the pixel. Local rather the pixel attributes can also be defined, either photometric (e.g. averaged intensity around $P$) or morphological (e.g. presence of a crossing of contours at $P$). The informational elements can bear a *linked* attribute, meaning that they are hyperlinks to other documents.

## Audio encoding for macro-analysis

In the macro-analysis phase, users can be either passive or active. In passive macro-analysis mode, the boundary locations and type of each object are transmitted to the user using vocal synthesis. The boundary location values correspond to a tactile Braille gradations surrounding the screen (see Figure 2). By using this gradation, the user can directly access a particular screen location.

**Figure 2: Tactile Braille gradations surrounding the screen.**

In active macro-analysis mode, the user explores the touch-sensitive screen; an auditory feedback is generated according to the type of touched element. Corresponding to the element type, different earcons $S(t)$ are used, and filtered with HRTFs RSX functions whose location parameters $\{\theta a , \theta e , d\}$ vary according to the screen location of the element.

## Audio encoding for micro-analysis

The micro–analysis phase lets the user actively focus on an informational element. If the selection consists of textual information, the simultaneous effect of having a finger on the text and depressing a given keyboard key initiates the reading of a word, a line, or a whole paragraph. Tonal variations are used to indicate hyperlinks.

If the selection occurs within an image, a composite waveform $S(t)$ is created as a function of the attributes of the touched pixel $P(x,y)$. In the current prototype and for pixel attributes, $S(t)$ is a function of either the luminance $L(P(x,y))$ or hue $h(P(x,y))$. The pixel position (x,y) in the image is transformed into location parameters $\{\theta a , \theta e , d\}$ in such a way that the virtual source $S(t)$ lies on a virtual vertical plane in front of the user. In this way, a mapping from the image to the virtual sound space is accomplished.

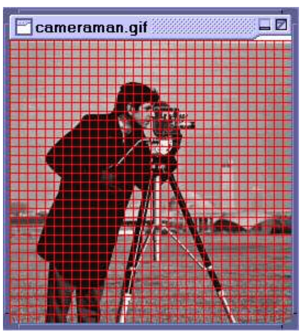


**Figure 3a: Bloc subdivision of the image.**

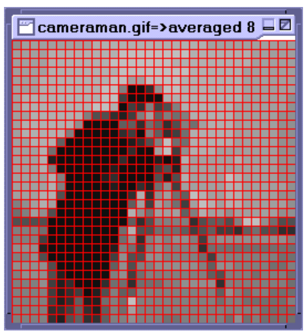


**Figure 3b: Display of the local attribute computed over each bloc.**

For implementing the mapping of local image attributes to sound space, the image is subdivided into square blocs (see Figure 3a). In the current implementation, *S(t)* is computed in each bloc as a function of the average grey-level of that bloc (Figure 3b).

## EVALUATION AND FUTURE DEVELOPMENTS

A first prototype using earcons for *S(t)* has been implemented and evaluated. Regarding macro-analysis, we have found that: (1) the sound space allows fairly good global layout determination, although the response delay due to the earcons stocking into the RSX buffer step is too long; (2) a passive mode in which the text-to-speech translator describes the screen content is useful. For micro-analysis, we have observed that: (3) the sound space is useful for helping the analysis of images, although shape recognition based on simple local image properties is difficult; (4) discrimination along the horizontal direction is easier than along the vertical direction. Finally, we have found that (5) the rendering of the elevation $\theta e$ is difficult. Insights gained, together with more results discrimination tests in virtual sound space, are being incorporated into a new prototype that uses more complex image to sound mappings. Also, for the sound space generation, we are trying to replace the RSX technology by directly programming the DSP and FM synthesiser chip of our sound card.

## ACKNOWLEDGMENTS

This project is financed by the Swiss Priority Program in Information and Communication Structures and by the Swiss Central Union Of and For the Blinds. The authors are grateful to A. Barrillier, P. Bovet, A. Bullinger and J. Conti, for their help in the design and evaluation of the prototype.

## REFERENCES

1. Blauert, J., Spatial Hearing, MIT Press, MA, 1983.

2. Brabyn, J.A., Developments in Electronic Aids for the Blind and visually Impaired, IEEE Eng. In Medicine and Biology Mag., Dec. 1985, pp. 33-37.

3. Cascading Style Sheets, W3C Recommendation, http://www.w3.org/TR/REC-CSS2, 1998.

4. Crispien, K., Würz, W., Weber, G., Using Spatial Audio for the Enhanced Presentation of Synthesised speech within Screen-Readers for Blind Computer Users, Computers for Handicapped Persons, Lect. Notes in Comp. Sc. 860, Springer-Verlag, 1994, pp. 144-153.

5. Intel Realistic Sound Experience, Intel, Inc., http://developper.intel.com/ial/rsx, 1998.

6. James, F., Presenting HTML in Audio: User Satisfaction with Audio Hypertext, ICAD 96 Proceedings, Nov. 1996,pp. 97-103.

7. Jaws for Windows, Henter Joyce, Inc., http://www.hj.com, 1998.

8. Kennel, A., Perrochon, L., Darvishi, A., WAB: World Wide Web Access for Blind and Visually Impaired Computer Users, in [6], pp. 297-306.

9. Kosslyn, S.M., Koenig, O., Wet Mind: The New Cognitive Neuroscience, The Free Press Macmillan, 1992.

10. New Technologies in the education of the Visually Handicapped, D. Burger, Ed., Les Editions INSERM, Paris, FR, Vol. 237, 1996.

11. Paciello, M.G., Hypermedia for people with Disabilities, ACM SIGLINK, V, 1, Febr. 1996, pp. 7-9.

12. Wenzel, E.M., Three-Dimensional Virtual Acoustic Displays, in Multimedia Interface Design, ACM Press, New York, 1992, Ch. 15, 257-288.

13. Winblatt, M., Browsing the World Wide Web in a Non-Visual Environment, ICAD 97 Proceedings, Nov. 1997.