Web site with recorded speech for visually impaired

Kenji Inoue[1], Toshihiko Tsujimoto[1], and Hirotake Nakashima[2]

[1] Graduate School of Information Science and Technology, [2] Department of Media Science,
Osaka Institute of Technology, 1-79-1 Kitayama, Hirakata City, Osaka, Japan
chikuwa.bushi@gmail.com, pa_o@hotmail.co.jp, nakas@is.oit.ac.jp

1 Introduction

Current assistive technology (such as a screen reader or voice browser) is not easy to use for the people who do not have good sight and/or who are not familiar with the computer. We therefore developed:

- A human-computer interaction model with the aural representation of a structural web page: the key idea is the **scanning findability** that can be achieved by properly shaped information in the aural representation.
- A web system that is **voice-enabled with recorded human voices** and works on top of the current common web browsers.
- **Three web sites of public service** built with this system, as an alternative to their original text-based contents to improve the information accessibility.

The Interaction Model with the Aural Representation of Web Pages

Web is easy to browse. Most people can immediately learn how to navigate themselves on the web, using the scroll bar and clicking links.

It is, however, not true for visually impaired. Visually impaired users need to use an assistive technology software such as a screen reader or voice browser. It needs bunch of key operations to go forward and backward, skip to the next link, jump to headings, follow "skip to content" link, change the frame, etc.

We do not want complicated key operations. Instead, we want to "scan" the page with a few fundamental operations. The key ideas to achieve that are:

- **Grasp of information**: recognition of the (vague) meanings of information should be processed unconsciously in the brain, rather not by an explicit operation consuming brain's attention resource.
- **Shape of information**: information objects themselves should show their distinct shapes to meaningfully state what they are and what they are of.
- **Scanning findability**: information or its clue the user want to know should be found by just scanning the page.

Aural Rendering vs. Visual Rendering

While we have no lack of examples for the visual representations of structured text objects, we have no consensus about their aural versions. Some candidates that will contribute to the scanning findability are as follows:

Table 1. Aural rendering methods vs. visual rendering methods

Aural rendering methods	Visual rendering methods
Volume adjustment	Font size
Vocal property (e.g. male/female)	Font color
Time margin	Margin or padding
Sound icon (earcon) or BGM	Icon or figure

2 Implemented System

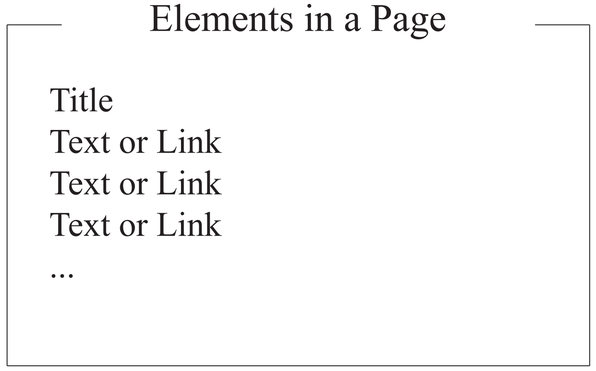
Our voice-enabled web system – we sometimes call it as "Voice Homepage" – is implemented as a **Java applet** or **Flash application**. Our objectives are as follows:

(1) To keep the number of types of operations minimum.
(2) To be integrated with the web technologies and no software installation needed where possible.
(3) To be able to use the recorded speech audio as well as the synthesized audio. The current desktop applications for visually impaired do not utilize much of the recorded human voices while the real-life systems such as navigation systems in stations seem to always use them. Usually recorded human voices offer higher quality and better recognition than synthesized voices.

Since the current web browsers do not natively support audio controls (i.e. CSS3 Speech Module), we built our system as a Java applet or Flash application.

3 Page Model

The information being presented with this system is first structured into a fundamental unit – a page. A page may have a title, some sentences, and/or links. They are modeled in the same manner as web pages, for example, a link is a directed link to another resource on the web. Browser history is supported as well.



4 User Interfaces

The output from the system is presented in the forms of both visual texts and audio. The font size of the texts can be changed.

The combination of the three types of audio can be used: recorded speech audio, text-to-speech audio, and sound icons. Note that the user interface softwares, composed as a Java applet or Flash application, do not have the ability to create the text-to-speech audio in real time on client side.

The main operations can be done with **4 arrow keys** (major key operations are listed in Table 2). Other keys may be used, but not necessary for browsing.

Table 2. Main defined key operations (excerpted)

Command	Key
Go and read next/previous text	Down / Up
Follow the current link	Right
Go back to the previous page	Left
Stop playing speech audio	Esc
Replay (re-read) the current text	r
Enlarge the text font size	+
Reduce the text font size	-

5 Applications

We have so far built up three web sites with this system. They are of public service: the aural versions of web sites for Hirakata NPO Center, Hirakata City, and Higashi Osaka City Fire Department. They offer the needed public information for their citizens.

Table 3 shows the basic statistics about the data sizes used in the above application web sites. The data size is the cumulative size used in the site and calculated from the .au Sun Audio files.

Each text is divided so as the audio representation to be less than 100 KB in size and 12 seconds in length, which would be the size that can be transmitted soon without waiting much.

The sound icons or aural rendering elements these sites are using include:

- Sound at the end of current reading speech
- Sound at the end of the page
- Male voice for texts and female voice for links

Table 3. Data size about the application sites

Name	Hirakata NPO Center	Hirakata City	Higashi Osaka City Fire Department
Number of Pages	183	78	80
Number of Recorded Audio Files	1701	443	443
Total Size of Recorded Audio Files	104 MB	24 MB	38 MB
Average Size of Audio	61 KB	54 MB	86 MB
Total Length of Recorded Audio (h:mm:ss)	3:47:24	1:23:58	0:52:47
Average Length of Recorded Audio	8.2 seconds	11.4 seconds	7.1 seconds