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

Abstract. Current assistive technology for visually impaired to access the web such as a screen reader or voice browser is not easy to use for the people who are not familiar with the computer. We investigate a human-computer interaction model with the aural representation of a structural web page. We then discuss about an overview of how scanning findability can be achieved by properly shaped information in the aural representation. We describe our web system that is voice-enabled with recorded human voices and works on top of the current common web browsers. We have so far built three web sites with this system, all of them are of public service, as an alternative to their original text-based contents to improve the information accessibility.

Keywords: Aural user interface, web accessibility, assistive technology, visually impaired

1 Introduction

Web is easy to browse. You look around a web page, navigating yourself with the scroll bar, and click a link if you find something that makes you interested. There might be some web sites hopelessly messed up so that you need to go back and forth across the web pages to find out where is what and what is where, and perhaps need to type some keywords in the search bar, but, basically, browsing is only of looking and clicking. That's all you need to do.

It is, however, not true for visually impaired who just cannot see the lights twinkling on the screen. They use the softwares of assistive technology such as a screen readers or voice browser. Both the screen reader and voice browser get the contents displayed on the screen and represent them to users with synthesized voice or braille output, but the voice browser reads texts itself while the screen reader handles the texts of other applications or operating system. In consequence, browsing with a screen reader needs complicated keyboard operations with modifier keys to avoid key collisions with the underlying applications being operated.

It gets better using a voice browser. No bunch of modifiers. Now you don't need to teach your grandma to be an Emacs user, but the typical total number of keyboard operation types doesn't change, as it reflects the complexity of the browsing model itself. Going forward and backward, skipping to the next link, jumping to headings, following "skip to contents" link, changing the frame, etc.

For sighted person they can be achieved by just clicking or dragging the scroll bar. Finding the next link, for example, done by your brain and not by an explicit operation. You can “scan” the page by looking around to find blue, underlined text object. The recognition is rather processed unconsciously, without consuming the valuable attention resource of the brain, than consciously, using the limited attention by planning and recalling the correspondence operations.

The key factor for that kind of scanning findability is that the information objects themselves show their distinct shapes to state what they are. Imagine the things written in a book, or just look this paper. Chapters are shown in large-sized font, with plenty of surrounding white space. A paragraph forms a rectangle-blocked region. List items hold a bullet mark in front of them. The navigation bar of the site, a example of the more higher-ordered semantic object, resides in the top (or left or right depending on the site) region of the site-wide consistent design.

While we have no lack of the visual representations of these structured text objects everywhere regardless whether it is physical or electronic, we have no consensus about their aural versions. Making use of the volume adjustment (vs. font size), vocal property such as male and female voices (vs. font color), time margin (vs. spacing), sound icon or earcon (vs. icon), or BGM would be the candidates for such aural rendering methods. It will ease the processing load of human brain, lessening the number of operations needed to be memorized.

This paper is, however, not for a holistic research to develop or investigate such aural presentation, but instead for a practicing web system, that is limited but actually deployed service for public welfare on behalf of visually impaired and other prospective users.

2 Implemented System

We developed a voice-enabled web system – we sometimes call it as "Voice Homepage” – which is implemented as a Java applet or Flash application, integrated with the current common web browsers and softwares.

Our direction of the system design was “easy to use” and whereby it can be used as many people as possible, and our objectives are as follows: (1) keeping the number of types of operations minimum, (2) integrated with the web technologies and no software installation needed where possible, and (3) recorded speech audio can be used as well as the synthesized audio. Since we have already discussed about the importance of minimizing the keyboard operation types, here we will talk about the integration and software installation and recorded speech.

Since this system is intended to be used widely, it would be better to only depend on the common environments, as otherwise it constructs an entry barrier. We, therefore, chose to use the common web browser to view our voice-enabled web sites. The current web browsers do not support audio controls, as they do not support the CSS3 Speech Module, so we built our system as a Java applet or Flash application, which are not the standard but most commonly installed hosting environments that work on top of the web browser. It makes that moving to and from other pages seamless.

Use of the recorded speech audio is notable in the context where the current desktop applications for visually impaired don't utilize much of the recorded human voices while the real-life systems such as navigation systems in stations seem to always use them. We all know that the recorded human voices offer higher quality and better recognition than synthesized voices. With this system we can use the recorded human voices, synthesized voices produced by text-to-speech, and sound icons mixed.

2.1 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, so, for example, a link is a directed link to another resource on the web.

2.2 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 for people who have low vision. 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, so the audio first need to be converted and deployed in certain audio file format that can be loaded by the system (.au for Java applet and .mp3 for Flash are the typical ones).

The input to the system is only by keyboard (but there is no reason not to support the pointing device in the future). The main operations can be done with 4 arrow keys: up and down to go forward and backward in the page, right to follow the link, and left to go back to the previous page before following the link. Other keys may be used, but not necessary for browsing. Other key operations are shown in Table 1.

Table 1. Main defined key operations (excerpted). More than one key may be defined to a command, but not listed here.

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	-
Restore the text font size to default	=

We encourage to use the recorded speech of human voices as far as possible. Synthesized speech does not produce recognizable high-quality audio, at least for the

next decade, as compared to, for example, the voices that are recorded by the trained announcers. Synthesized voices may also need the users to be trained to get used to the pronunciations to recognize what the sounds are saying. Since the quality of the synthesized voices depends on the language and the developed technologies for the language, how long it takes may differ in the languages.

Sound icons, or earcons, can be used to construct the shapes of information as well as the system feedback to the user's operation. A sound at the end of the current reading speech can let users to move to the next text. A sound at the end of the page lets users know where they are now at. A deep sound after entering a page may indicate it is far away from the home page of the site. Usual texts and links may have different sound icons. It is leaved to the aural designers.

Changing the playing speed of audio, without changing its pitch, should have been implemented, but the current system could not support it due to the technological issues. It is one of the key feature to make the scanning of the pages possible, in combination with the properly aurally-rendered representation.

3 Applications

We have so far built up three web sites with this system. They are of public services: 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. For example, Higashi Osaka City Fire Department offers the information such as how to call an ambulance, what we should do when a disaster happened, and how to give the first aid. Such information in audio would also be helpful for the foreign citizens who can speak the national language but cannot read or write it.

Table 2 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. For the site of Hirakata NPO Center, almost 4 hours of speech was recorded, and much more time was taken to process them to use them in the system. It shows that it takes certain amount of efforts are required in addition to preparing the texts.

Table 2. Data Sizes 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 KB	86 KB
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

4 Conclusion

First, we have modeled the structural information objects in the web pages to have their aural shapes of representation to make it easy to scan and find the information the users need by lessening the processing load of recognition on the brain.

Second, we have built a half-experimental but practically usable web system that is easy to use for visually impaired. The system makes use of the combination of the recorded speech, synthesized speech, and sound icons.

Third, we have created three actual applications of this system. All of them are of public service and they utilized this system as an alternative to their original text-based web contents to improve the information accessibility.

There were many pros and cons from visually impaired for this approach, but we have not yet conducted a systematic evaluation of the system. It remains as an issue and we need to further investigate and improve the system and its model.

Acknowledgements. We thank to all the staff at Hirakata NPO Center, Hirakata City, Higashi Osaka City Fire Department, our lab students and voluntary people who made this research possible. I believe that you all are the greatest people increasing the amount of happiness, if such exists.