WEB-SPEAK: A CUSTOMIZABLE SPEECH-BASED WEB NAVIGATION INTERFACE FOR PEOPLE WITH DISABILITIES

Foad Hamidi
CanAssist
University of Victoria, Victoria, BC, Canada
foad@canassist.ca

Leo Spalteholz
Department of Electrical and Computer Engineering University of Victoria, BC, Canada
leos@ece.uvic.ca

Nigel Livingston
CanAssist
University of Victoria, Victoria, BC, Canada
njl@uvic.ca

1. INTRODUCTION

In recent years, the increasing importance of access to the World Wide Web has necessitated the development of alternative technologies that can facilitate Internet access for people with disabilities. In previous research, we proposed a new system that allows users unable to accurately operate a pointing device (such as a mouse) to efficiently navigate the web. In that system, users could select any element (links, buttons, form fields) on a web page by typing one or two characters [4]. The web navigation system was designed to minimize the number of required characters to uniquely select any element, and to separate the typing interface from the selection system. This allowed the user to navigate the web using any input device that was capable of producing alphanumeric characters.

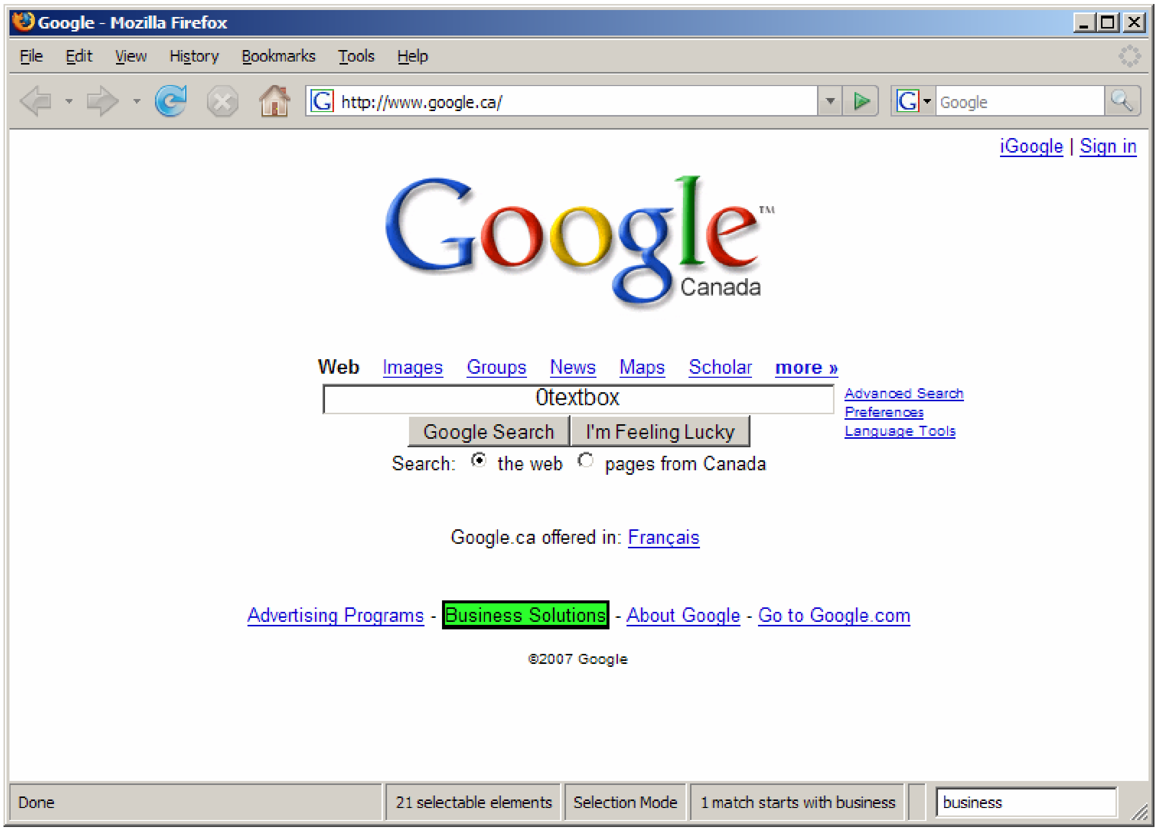
In this work, we take advantage of these properties to develop an application, *Web-Speak,* which - by integrating a speech recognition interface into the web navigation facility - allows the user to navigate the web using a limited number of verbal commands and keywords. We argue that this approach provides users who can not use a mouse, keyboard, or other haptic input devices with a simple alternative interface that is highly customizable and robust. Furthermore, by only requiring a small set of customizable commands, the system is suitable for people with speech impediments or non-native English speakers who can not use standard voice recognition engines.

2. WEB-SPEAK: A SPEECH-BASED INTERFACE FOR WEB NAVIGATION

Our speech-driven web navigation application, *Web-Speak,* is developed as an extension to the open source Mozilla Firefox web browser. Our speech recognition component is built on the open source, Java-based Sphinx-4 speech recognition engine [5].

In our previously proposed method, in order to navigate to a link on a webpage, the user needed to input very few keystrokes [4]. This method was implemented by assigning labels to all the selectable elements in a loaded web page. When the user entered a letter, an algorithm searched the labels assigned to the page elements to find the element that the user most likely wanted to select. Matching elements was done by taking several factors of the elements into account, such as visual prominence, their position on the page, and the context around the match. Once the most likely match for the input letters was identified, it was highlighted and could be selected by pressing the “enter” key.

Web-Speak uses speech as input to the web navigation system. When a page loads, the Firefox extension compiles a list of all the distinct words present within link text on a page and transmits this list to the speech recognition component. *Web-Speak* receives the list of available words, and can then work in two modes: *select* and *spell*. In the *select* mode, it restricts recognition to only the list of available words, which vastly reduces the complexity of recognizing a command. Constraining the operating set in this way can make recognition robust even without training. To select a link, the user may say any word that is present in its text. After recognizing a word, the application will send the text to the browser, which will use the matching algorithm to find and highlight all the elements on the webpage that include the recognized word. After saying enough words to uniquely identify the desired element on the page, the user can select that element using a special command keyword. Figure 1 shows the Google home page after the word "Business" is uttered.


Google - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.google.ca/
Google
iGoogle | Sign in
Google Canada
Web Images Groups News Maps Scholar more »
0textbox
Advanced Search
Preferences
Language Tools
Google Search
I'm Feeling Lucky
Search: ● the web ○ pages from Canada
Google.ca offered in: Français
Advertising Programs - Business Solutions - About Google - Go to Google.com
©2007 Google

Done | 21 selectable elements | Selection Mode | 1 match starts with business | business

Figure 1: The Google home page after the word "business" is uttered. If the keyword "yes" is spoken the green link will be followed.

The advantage of the *select* mode is that the selection method is very direct and does not require the user to memorize any keywords (other than the system navigation commands that can be optionally displayed on a tool tip box). A similar web navigation approach has recently been implemented in the popular consumer voice recognition package, Dragon Naturally Speaking [3].

While the *select* mode is suitable for native-English speakers with clear speech, it might cause hindrance for users whose speech is limited or have a strong accent [1]. Although the set of words in the *select* mode is usually small, this set is based on the page content, and could contain any word present on the page, including words that the speech engine cannot recognize, or that the user will have difficulty pronouncing. To address these problems, we have developed an alternate usage mode.

The *spell* mode, as the name implies, involves the generation of single letters and digits one at a time. Spelling using speech recognition software is problematic because the amount of information contained in the pronunciation of similar sounding letters is often not enough to differentiate between them (e.g. “m” and “n”) [2].

To overcome this problem, in an approach similar to the usage of the International Phonetic Alphabet (IPA), we associate each letter of the alphabet with a *keyword* starting with that letter. This approach makes speech recognition very robust, because, in addition to having a very small vocabulary in the speech recognition engine, the keywords provide much more auditory information than individual letters.

As the IPA was designed to facilitate recognition of words over poor quality communication channels, we initially adopted it for the set of keywords representing the English alphabet. We conducted a small scale user study to determine untrained recognition accuracy for a set of 47 keywords consisting of the IPA words, numbers and a few commands[1] involving 17 subjects of varying age (20 to 60 years), gender, and ethnic backgrounds. Subjects were prompted to say a word, and the engine attempted to recognize their response. Each subject was prompted for each keyword in the alphabet twice, in random order, for a total of 94 prompts. Although results were quite good (average 92.7% accuracy on first attempt), we identified several problem words that the Sphinx engine consistently had trouble recognizing.[2] To improve accuracy, we replaced problem words with alternatives and retested on a different group of 15 participants. The average recognition accuracy improved with our modified keywords, to 95.8%.

The set of keywords used by Web-Speak can be customized to maximize recognition accuracy or appeal for a particular user or user group. For example, we have created an alternative alphabet for children that consist of animal names. Also, keywords can be modified if a user has pronunciation difficulties with the standard word. We speculate that while initially the user must learn the keywords, with frequent use the effort of remembering the words will become very small. In the example shown in Figure 1, to select the “Business solutions” link, the user only has to speak the keyword associated with the letter “b” (e.g. “Bravo”).

3. CONCLUSIONS

[1] The commands consisted of “yes”, “delete”, “select”, “spell”, “space”, “scroll up”, “down”, “back”, “forward”, “favorite” and “next”.

[2] The set of problematic words consisted of “papa”, “echo”, “Quebec”, “victor” and “kilo” which were replaced by “piano”, “elephant”, “question”, “Victoria” and “kingdom”.

We have presented an application that integrates a speech recognition interface into our previously developed web navigation system. Our application, *Web-Speak,* works in two modes. In the *select* mode, suitable for native English speakers with clear speech, the user selects links and other web page elements by saying any word contained in the text of the desired link. The *spell* mode, in contrast, only requires users to be able to clearly pronounce a very limited number of keywords that are associated with letters and system commands. This mode not only makes the application robust and accessible for users with limited speech or strong accents, it also allows for high customizability as suitable keywords can be assigned for each user to minimize recognition errors and ease of usage.

4. REFERENCES

[1] Derwing, T. M., Munro, M. J. and Carbonaro M. Does Popular Speech Recognition Software Work with ESL Speech? *TESOL Quarterly*, Vol. 34, No. 3, TESOL in the 21st Century. (Autumn, 2000), pp. 592-603.

[2] Marx, M. and Schmandt, C. Putting people first: specifying proper names in speech interfaces. In *Proceedings of the 7th Annual ACM Symposium on User interface Software and Technology* (Marina del Rey, California, United States, November 02 - 04, 1994). UIST '94. ACM Press, New York, NY, 29-37.

[3] Nuance. Dragon NaturallySpeaking 9. http://www.nuance.com/naturallyspeaking/, 2004.

[4] Spalteholz, L., Li, K. F., and Livingston, N. Efficient navigation on the world wide web for the physically disabled. In *proceedings of the 3rd International Conference on Web Information Systems and Technologies,* pages 321-326, Mar. 3-6, 2007.

[5] Walker, W., Lamere, P., Kwok, P., Raj B., Singh R., Gouvea, E., Wolf, P. and Woelfel, J. Sphinx-4: A flexible open source framework for speech recognition, Tech. Rep., Sun Microsystems Inc., 2004.