AN INVESTIGATION INTO VOICE CONTROLLED WEB BROWSING FOR THE ELDERLY: VOICE CONTROLLED INTERFACE

Cole D. Noble

School of Electrical & Information Engineering, University of the Witwatersrand, Private Bag 3, 2050, Johannesburg, South Africa

**Abstract:**

US statistics for 2010 provide indication that the elderly (those over the age of 65) are the fastest growing group of new internet users in the country. Also, even though they comprise just 14% of the population, they only contribute to 4% of the national internet usage. These stats give a broader indication regarding the need for computers to be usable for the elderly. This paper seeks to provide a platform for improving the interface of computers, specifically the internet and web browsing, using voice recognition and basic forms of feedback. Three voice enabled WebPages were made to carry out tests on the elderly using two different techniques of voice referencing and three different feedback techniques. The two voice referencing techniques include: Spoken Numerical Referencing of Links (Numerical Referencing) and Spoken Link Name Referencing (Link Name Referencing). The first method assigns numbers to every link on a page and allows individuals to reference every link by its number. The second method highlights specific words in links and allows users to access links by speaking those names. Three forms of feedback were also tried, namely, pop-ups, link highlighting and audio feedback. The two voice referencing methods were tested for comparative performance and preference while the feedback methods were tested briefly to simply gauge user preference. The conducted tests indicate superior performance of Link Name Referencing over Numerical Referencing on simple pages and the opposite being true on more complicated pages. Feedback preference was indicated to be link highlighting followed by verbal feedback. All implementation was conducted using an online speech API, *JavaScript* and *HTML*. Future work could include creating a generic web browser add-on or other more complicated combinations of visual annotations and voice referencing techniques.

**Key words:** API, elderly, internet, links, referencing

# Introduction

The following paper presents the problem of internet usability for the elderly, and attempts to begin to address the problem by investigating two voice recognition techniques and a few visual feedback techniques.

The paper starts by explaining the need to improve the usability of the internet for the elderly. It then narrows the problem down to a set of questions pertaining to the elderly and different potential forms of interface techniques. Specific requirements for different software iterations undertaken to answer such questions are then outlined. This is followed by a brief mention of the limiting factors to the investigation and a few assumptions that were made before proceeding with the investigation. The exact procedures undertaken for each of the 3 iterations and some ethical considerations made before commencing the tests are then provided. This is followed by a summary of the main results observed from the investigation. An implementation overview is next provided which is followed by some of the problems encountered in the project.

The tools used to perform the tests as well as some aspects of group work are then mentioned. Following this, a time breakdown is referenced. Lastly a few critiques of the project and a few future recommendations for investigation are given.

# Problem background

Statistics from the US indicate that those over the age of 65 comprise the fastest growing group of new internet users in the country (see Appendix B). 2010 stats indicate that the US population is comprised of 13% of the elderly and that only 4% of the total internet usage is contributed to by this group.

This lack of usage of the internet can, in part, be attributed to two factors. One: A lack of computer literacy in the elder generation. Two: Deterioration of health may inhibit the use of interfaces required for the use of computers.

With time the first factor is expected to change as more and more computer literate individual’s age. The second factor, health deterioration, could result from many health related problems. These problems include, but are not limited to, limited use of hands due to arthritis or carpel tunnel syndrome, eyesight deterioration and hearing deterioration.

The elderly, at least in the US, are said to have more disposable income and time at their disposal. This further increases the need for the accessibility of online entertainment, shopping, banking and communication for elderly individuals.

From the points mentioned above it can be seen that this is a largely unreached group of potential internet/computer users in a significant first world country. Also, as one of the reasons for the lack of penetration into this ‘market’ will naturally change with time (computer literacy), health limiting factors need to be overcome. Voice recognition is one way of circumventing the issue of impaired hand coordination and various forms of visual feedback could be incorporated into applications to mitigate the inhibiting effects of diminished eyesight and hearing.

Voice recognition is unfortunately less accurate on the elderly. Research however indicates that creating special acoustic models for the elderly can improve the performance of voice recognition by up to 14%.

# General problem

As mentioned in Section 2 above, voice recognition is one way of working around the problem of impaired hand usage. The question thus arises, what sort of voice referencing techniques can be used to improve the usability of computers for the elderly. As a secondary question, feedback techniques that may be more suited to visual and hearing impairments were investigation as well as any additional requirements that may arise for the usage of computers for the elderly.

# Existing solutions

Appendix B provides a more extensive description of existing solutions. There are some key points from existing solutions that will however now be highlighted to the extent that they informed the current implementation of the project.

Research conducted broke down existing solutions into three categories, namely, *Voice Recognition software*, *Visual Rendering* and *Combined System*. The first dealt with web browser relevant software that would explicitly perform voice recognition. This took the form of either add-ons (such as *VoiceBrowsing* for *Internet Explorer*) or special browsers (such as *Opera*). *Visual Rendering* explored packages that provided forms of numerical annotations to web pages specifically (such as the *Mozilla Firefox* add-on, *Mouseless Browsing* or *Pentadactyl).* A third kind of package was investigated, namely, a kind that incorporated elements from the first two systems. Only one of these systems was investigated, namely, *Dragon NaturallySpeaking*. This software is general purpose speech recognition software that lets you navigate your computer or browse the web using speech recognition. It incorporates both forms of referencing compared in this investigation (see Section 6.1) by letting you speak whatever you want to go to and numbering choices only if there is conflict occurring.

# Specific Questions

Section XXX highlights the broad question and purpose of the project. In light of existing relevant applications mentioned above, the investigation specifically sought to combine *Visual Rendering* (numbering of links) with *Voice Recognition Software* in a different way to the *Combined System* mentioned above. Instead of creating a hybrid been numbers and names, two isolated techniques were considered for testing. Firstly, assigning every link a number and allowing users to speak those numbers to navigate (Numerical Referencing). Secondly, highlighting specific link words and allowing users to speak those words to reference a link (Link Name Referencing). With these two methods in mind, and the general problem provided in Section 3, the following questions arise.

Can Word Error Rates (WER’s) of voice recognition be reduced without developing a special acoustic model for the elderly by limiting the word space to short, more commonly pronounced, numbers?

Is the use of numerical visual annotations a preferred way of referencing elements on a web page for the elderly?

What sort of feedback techniques (pop-ups, link highlighting or voice feedback) provide a preferred form of feedback for the elderly?

What areas of voice navigation for the elderly require further investigation?

# Requirements for answering SPECIFIC questions

## Requirements overview

In order to answer the above-mentioned questions, three implementations of various voice enabled websites were used. Each of these was termed an “iteration” of the investigation. Each iteration had specific goals/requirements that were pencilled just prior to the implementation of each iteration.

Success criteria will mean deriving an answer to each of the following iteration goals.

## Iteration 1

* Determine the relative performance of numerical referencing over spoken link names
* Determine the preference over the two techniques
* Determine what sort of feedback is preferred (pop-ups, highlighting, voice feedback)

## Iteration 2

* Isolate comparison between performance of numerical and spoken link names better by introducing training session

## Iteration 3

* determine whether the performance and preference of the two techniques persists on more complicated websites
* Determine additional feedback on the feedback techniques one implemented
* Introduce additional techniques to try and improve the browsing experience and test their applicability (confirmations)

# Limiting factors

A few factors constrained the implementation of the various investigation iterations.

Firstly, cost was a big factor to consider for the project implementation. This constrained the development tools to being *freeware* (see section XXX). A total budget of R600 was allocated for the project. This money was used in part to purchase: data-bundles required for internet access when testing implementation, microphone headsets and printing a poster for project demonstration.

Time was another constraint on the project. A total of 8 weeks was allocated to the project. Within this time, implementation had to be tested for each iteration and adequate testing needed to be conducted. Each test required time consuming training before each test. This required additional time allocation.

# Assumptions Made

All the test subjects were assumed to not be computer literate. This meant that earlier tests had to be very sequential in nature and require little knowledge of computers.

It was assumed that subjects would have adequate sensory perception and mental competence to conduct tests and provide suitable feedback.

Although elderly are generally deemed as those over the age of 65 (See Appendix B), the elderly for this investigation were assumed to be those over the age of 55. This aided with the number of potential test subjects.

# Investigation procedure

## Procedure introduction

The three iterations mentioned and in section XXX, were carried out to achieve the mentioned requirements. Each of these will now be explained in their content. Please note that the surveys for these iterations are given in Appendix XXX.

## Iteration 1

The first iteration was derived to perform a comparison between the two mentioned referencing techniques (numerical and spoken link name). It was also constructed to determine a preference of feedback (pop-ups, link highlighting and verbal feedback).

The test consisted of a series of questions that required the user to navigate through two links to select the answer for each question. Interesting animal fact questions were used so as not to confuse individuals and stimulate some degree of interest in the test. The users were required to conduct two tests with numerical referencing and two tests with Spoken Link Names. Three remaining sets of questions (each also consisting of two questions each) where used to expose users to pop-up messages, link highlighting, and voice feedback respectively. All errors, being any misinterpretation by the Speech API, were recorded and counted for each question.

## Iteration 2

The second iteration was designed to correct the possible inconsistencies of the results in the first iteration (see section XXX), the theory was that since numerical referencing was used as the first method each user encountered, the results reflected a time when users were still getting well acquainted with the system. Furthermore, since the numerical referencing system was also used to show the users the different feedback techniques, it was felt that they were more inclined to prefer numerical referencing.

For these reasons, the second iteration differed from the first iteration in the following aspects:

* Only the two referencing techniques were tested (no feedback techniques). It was felt that the first test gave an adequate indication of user preference.
* The number of questions for each technique was increased from 2 to 5 questions. This was felt to be necessary so that results gave a better statistical average.
* Two training questions were introduced for each referencing style. This was so that the users had some time to get accustomed to each referencing style.

## Iteration 3

The third iteration was implemented to employ a few of the findings from the earlier tests. This was undertaken for various reasons:

* For one, to get a qualitative indication of how users now experienced the two types of referencing styles on a more complicated page.
* To get a measureable indication of the recognition errors resulting from the two methods. This will help determine if the two referencing methods still perform in the ways indicated by prior iteration results on pages with less links.
* Implement an additional confirmation feature that allows users to confirm the Speech APIs interpretation before a command is carried out.

In this test, a facsimile of a news website was created. A screenshot of this website can be seen in appendix XXX. This website employed both of the techniques separately. Users were then required to perform certain use-cases on the website (such as navigating, scrolling etc.). Error rates were recorded and miscellaneous questions were posed.

See Appendix XXX for the survey, Section XXX for result write up and Appendix XXX for the results graphs.

# Ethical considerations

A few ethical considerations come into play with the testing of this project, most of which concerned the test subjects.

Some of the subjects tested were fairly frail and had medical conditions that were significant and required due attention. One lady, for instance, had a heart condition. To make sure that test subjects did not feel pressurised in any way (so that medical conditions were not aggravated), subjects were clearly told that if they felt uncomfortable at any stage of the test they were not obligated to continue. Furthermore, test subjects were guided through the program and where not expected to have to navigate towards the correct answer on their own (they were often told what to say).

To further reduce the strain on the subjects, the number of maximum repetitions following recognition errors was limited to 3. That is, if people kept on making mistakes with certain numbers or words, they were allowed to move on to the next question in the test after 3 tries.

The Speech API is licence under GNU General Public License [1]. This implies that all versions of the project, so long as it uses this licence, need to remain under the same licence.

# Results

## Primary results

The results found from the various iterations are given in appendix XXX. The main results are mentioned here:

In Appendix XXX, Figure 2, a graph is given indicating user preference for feedback. As can be seen, most people preferred having visual feedback in the form of link highlighting. Since there is a contingent of users who prefer having audio feedback, the 3rd iteration of the project incorporated these two preferred feedback techniques in a compound way.

Appendix XXX, Figure 2, gives a graph of user referencing preference between the iterations. It is evident that the first 2 iterations show a preference for numerical referencing while for the 3rd iteration, there is no glaring preference for either technique. The 2nd Iteration bears a very close resembles to the 1st Iteration (see section XX). It was thus expected that the results be similar for user preference. Since this is indeed the case, it shows that the elderly do prefer numerical referencing on simpler web pages. The preference for the 3rd iteration shows that, on more complicated pages, the sample preferred neither referencing technique.

Figure 3, in Appendix XXX, shows the relative errors experienced by each of the referencing techniques. As can be seen from the graph, the first Iterations saw an equal relative number of errors. This indicates two things: For one, it shows a consistency in results across the two iterations. Secondly, it indicates that on very basic web pages (those with few links), numerical referencing seems to perform worse. Also to consider from these results, is that Numerical Referencing experienced marginally better results in the third Iteration, which uses more complex pages.

## Additionally derived results

Some other graphs were plotted from the extracted results. This was done to try and confirm suspicions about the nature of age and gender performance with speech recognition.

With regard to age, the data provided in Appendix B indicates that performance of speech recognition should deteriorate with age. Figure 4 in Appendix XX gives a graph plotting the average errors per individual in each age group. As can be seen, generally, there is an increase in the error rate with age. There is one exception though, being the error rate of 75-85 year olds in the 3rd Iteration. Only two people were tested from this group and these people spoke exceptionally well.

Lastly, there was an expected higher error rate for males over females (Appendix B). Figure 5 in Appendix XX illustrates the average error rate per individual in each age group. The graph seems to indicate a contrary trend to what stats expect. That is, females experienced more errors on average than males. This result is also not considered to be a good indication of the general trend because there were significantly more females tested than males. Thus these results are considered results of exceptional individuals rather than a general trend. More testing would be required to confirm this.

A few other results were also derived from questions posed during the iterations. See appendix XXX.

# Implementation overview

The developed code essentially evolved through the three iterations of the project. For this reason, the implementation will be discussed in light of the most complex iteration, Iteration 3. Most of the functionality remaining in Iteration 3 could be used in earlier iterations as well, provided that the appropriate parameters are set. However, since no more changes were required in earlier iterations of the project as iterations progressed, the *JavaScript* code used for each iteration was separated for each iteration.

An overview of the flow of the web application is given in Figure XXX in Appendix XX. Each web page includes various segments of JavaScript. This code serves to provide an interface to the flash component used to stream audio to and from the server and to process the results from the server. The overview of the processes involved will now be discussed in brief.

Once the JavaScript has been loaded, the necessary flash component is set up (see section XXX). The page is then processed according to the mode set (either Numerical or Spoken Name) and the vocabulary required for a certain page is determined. This vocabulary then gets initialised onto the server so that the server knows what words to expect. From here, the program waits for the voice recognition to be activated through the flash component. When this occurs, that is when the *Ctrl* button is depressed, the user’s voice input from the microphone is then streamed to the server through the flash component until a stop recognition command is given (until *Ctrl* button is released). The program will then wait for the server to return a result (either a known word or the words, “recognition error”). The result is then processed and classified according to the type of command being issued.

If the result could potentially lead to navigation, a confirmation of whether the command was recognised correctly is required (see section XXX) before navigation is carried out. Confirmation, in the programs context, means that the user must either say, “Yes” or “No”. If confirmation mode (*confirmationMode* global) is set to true and the user confirms positively, a link is followed or the command is performed, otherwise the code will refresh the pages.

If the result pertains to link navigation, the link is highlighted, verbal feedback is given and the confirmation mode is changed. Otherwise, if the command is not associated with link navigation, the user will receive verbal feedback from the program indicating choice and the confirmation mode will also be changed. Once confirmation mode has been set, the vocabulary on the server is set to only, “Yes” and “No”. The user will then need to depress and speak in to the microphone to confirm or decline the speech recognition’s interpretation of what is said. If what is said this time is an affirmation of the command (“yes”), then the highlighted link is followed or specific command is carried out. Otherwise, if no, the highlighting of the link is undone (if it was a link type command), and the vocabulary is set up again for the page.

On the other hand, if the result is of a type that does not require confirmation (namely, “up” to scroll up or “down” to scroll down) then the command is simply executed. Please see Appendix XXX for more implementation details.

# Problems encountered

The main problem encountered was the reliance of the system on an online API. On a couple of occasions the server went down. It took a little time to realise this and so some time was squandered trying to fix problems that were never actually there. Communication was established with one of the developers who would reset the server every time we experienced a downtime. However, unfortunately at *Open Day*, the server went down. Since the server is in a different time-zone, the lag between communications meant that the server could not be reset by the end of the day. Fortunately, a demonstration video had been constructed for such an eventuality.

Other problems arose with *Eclipse*. Often, unpredictable behaviour would occur. This was avoided by using the text editor mentioned in section XXX.

# Tools

As discussed in section xxx, one of the constraints of conducting the investigation was conforming to cost requirements. For this reason free tools had to be used. Initially, Eclipse [2] was used as the IDE of choice to write the JavaScript and HTML. However, Eclipse also ran slowly on the development computers and had unpredictable behaviour. Since JavaScript is a loosely typed language, the usual advantages of IDE’s are not really experienced. Instead, a free text editor with syntax highlighting was used. Namely, *Notepad++* [3] to do syntax checking. This proved much faster and easier to use. *Firebug* [4] for *Firefox* was used to debug the syntax.

Importantly, a free, existing API, SpeechAPI, was used to perform all the voice processing [5].

The webpage was hosted locally using IIS [6]. Running the webpage locally meant that there once was one less point of reliance on an externally hosted server.

*Git* was used to monitor version control [7]. *Github* was used to host the repository [8]. This helped tremendously with merging and monitoring changes.

Toggl [9]was used to keep track of the time spent on the project. The total record of time spent on the project was XXX hours. See appendix XXX.

# Group work and division

A breakdown of work division is provided in Appendix A. As was mentioned in section XXX, *Git* and *Github* were used to monitor and aid work cohesion. Some of these graphs are also provided in the mentioned appendix.

# Critical analysis

## Critique of investigation

Sample size and sample subjects may have played a significant role in the overall results experienced. A total of 18 people were tested (See Appendix XXX). A greater effort should have been put into ensuring an even division amongst age groups, gender and age. Ideally, the sample size should have been bigger to compensate for the statistical anomalies. However, this experiment was not conducted to gain firm statistical proof of technique performance and preference. It was merely instituted as a starting point for determining what direction such an investigation should take (what areas need to be looked into further).

Another aspect of the investigation is the usage of the specific voice recognition API. No knowledge is had as to the methods employed in the voice recognition on the server-side. Thus, these results may just give an indication of performance of elderly voices for this particular API, and not necessarily an indication of performance with general voice recognition software. However, this factor does not excuse the preference factor for the elderly (between numerical referencing and spoken name referencing or between the feedback techniques). We are thus still able to conclude that there is a general preference for numerical referencing on more simple web sites.

## Tradeoffs

Since the investigation was set out primarily as an assessment revolving around the elderly, it was necessary that the first Iteration of the project could be rolled out with sufficient time to conduct the required tests.

Initially, the system was proposed to be a generic *Firefox* add-on which could add the required voice functionality to most WebPages (See Appendix B). However, since all that was required was to be able to conduct a few tests, the use of voice enabled WebPages was chosen in contrast to the generic solution. See section XXX for future recommendations.

This time constraint also meant that a relatively easy method of achieving voice recognition on a website needed to be found. For this reason the chosen voice API was used. This API however may not have been the best option with regard to performance and furthermore, the lack of knowledge pertaining to the API’s implementation means that results may not necessarily be applicable to all speech recognition engines. As mentioned in Section XXX, this method still however enabled a clear enough idea of user preference for the main two methods of voice referencing and the methods of voice feedback. Another down side of using this API was the reliance on the server. This meant that whenever the server went down, the program was useless. Fortunately, this never occurred with testing and testing is often flexible enough such that rescheduling is possible in such an event.

## Recommendations

This investigation was conducted to serve as a starting point for solving the questions presented in Section XXX. For this reason, and in light of the experienced results, there are many avenues for further investigation.

A more generic way of highlighting words in links could be investigated. For instance, instead of hard-coding every link with a key-word ID, the page could process every link, use some sort of means to determine which words are key in meaning and which words are sufficiently different for adequate recognition performance.

A client side application/browser-add-on could be created to generically annotate WebPages. This application could link with an online API which provides an indication of tolerance. This could be used to determine whether or not confirmations are required for specific links. . An API which provides such a service is XXX [10]. Unfortunately, this particular API is not free, so further investigation may be required or the possibility of using client side voice recognition instead.

If the usage of an online API is continued, then a means of automatically detecting a spoken command needs to be investigated. The currently used API does have an automatic mode, but its accuracy significantly decreases with this mode. The possibility of developing a device that can easily be used to activate recognition could be investigated (for instance, a hand held device that the elderly can depress to activate recognition).

Since numerical referencing seems to be preferred for smaller amounts of content, a better means of breaking down more complicated websites could be investigated. For instance, sections could be referred to by a name, if a section is small enough, those items within sections can then be numerically annotated. In this way, the numbers used for numerical annotation need not grow too large as each section may have its own new set of numbers. Number annotations could also vary according to the existing methods of number overlaying (see Appendix B). For instance, numbers could be reassigned according to every view as is used in XXX, or they could appear temporarily above links when a certain command is given as in XXX.

Colour calibration options could be included so that individuals can have links shown in specific colours or be highlighted in specific ways. Furthermore, an investigation into section highlighting could be conducted to see if different section colouration makes it easier to understand/read a webpage. Various techniques could also be looked into to determine how one could allow elderly users to zoom into different web page sections using voice (so possibly overlaying lines indicating webpage sections that are ‘zoom-referable’).

# Conclusion

The issues surrounding the quickly growing group of new internet users, the elderly, were discussed. One of the problems with this group of internet users was seen to be health issues that arise, specifically problems with hands, eyes and ears. Voice recognition was proposed as one of the ways of improving the usability of computers for the elderly and thus circumventing the health issues related to hands.

Voice recognition was incorporated into 3 specially created websites which were combined with various forms of visual feedback. The purpose of these websites was to test what voice recognition techniques and forms of visual feedback could improve the usage of the internet for the elderly. The learned techniques could potentially be used to improve the usability of computers for the elderly.

The voice recognition functionality was achieved by using an online API. Further functionality was attained using *JavaScript* and *HTML*.

Each website was termed an “iteration”. Tests were then carried out sequentially on the elderly (defined as 55-85 years old in this paper) for each iteration. The elderly were tested so as not to aggravate their health problems. Various results were found: The main results being that numerical referencing is preferred for more simple pages to spoken link name referencing, but spoken link name referring out-performs the numerical referencing technique. On more complicated websites, the preference difference between the two techniques is negligible, but numerical referencing out-performs the other method. Visual preferences were found to be link highlighting and voice feedback.

Tests should potentially have been carried out on more individuals with an even distribution of age and gender. A time limitation meant that this was not possible.

This investigation was not intended to be complete in nature. Rather, to serve as a starting block for the presented topic. See section XXX for more details.

# Bibliography

1. **Foundation, Free Software.** GNU General Public License. *GNU Operating System.* [Online] 2007. [Cited: 26 October 2011.] http://www.gnu.org/licenses/gpl.html.

2. **Eclipse Foundation.** Announcements and News. *Eclipse.* [Online] 2011. [Cited: 26 October 2011.] http://www.eclipse.org/.

3. **Ho, Don.** About. *Notepad++.* [Online] 2011. [Cited: 26 October 2011.] http://notepad-plus-plus.org/.

4. **Mozzila.** Firebuf -Web Development Evolved. [Online] 2010. [Cited: 26 October 2011.] http://getfirebug.com/.

5. **Lord, Spencer.** Online Speech Recognition API. *SpeechAPI.* [Online] 2010. [Cited: 26 October 2011.] http://www.speechapi.com/.

6. **Microsoft Corporation.** IIS OVerview. *IIS.* [Online] 2011. [Cited: 26 October 2011.] http://www.iis.net/overview.

7. **Chacon, Scott.** Git -the fast version control system. *Git is.* [Online] 2011. [Cited: 26 October 2011.] http://git-scm.com/.

8. **GitHub Inc.** Github Social Coding. [Online] 2011. [Cited: 26 October 2011.] http://git-scm.com/.

9. **Toggl.** Tasks. *Toggl.* [Online] 2011. [Cited: 26 October 2011.] https://www.toggl.com/tasks.

10. **ISpeech.** iSpeech API Specification Version 2.0. *ispeech.* [Online] 2011. [Cited: 26 October 2011.] http://www.ispeech.org/api.