Voice Navigation of Structured Web Spaces

Dr. Manuel A. Pérez-Quiñones <perez@cs.vt.edu>
Natasha Dannenberg <ndannenb@vt.edu>
Robert Capra <rcapra@vt.edu>

Department of Computer Science
Virginia Tech
Blacksburg, VA 24073

Abstract
Voice Navigation of web spaces has become a reality in the last few years, partly due to the rapid adoption of VoiceXML and the increase in communication quality and computing power of cell-phones. This report discusses some of the approaches on how to convert the web content to a way that could be used from a voice-enabled phone. These approaches have different pros and cons when it comes to the usability of the voice-navigation of web spaces. This report discusses our work to produce voice navigation of web spaces placing usability as the highest criteria. We have designed of a voice user interface for pages with a fixed structure and user-specified content, such as My Yahoo! pages. For these types of pages, we have defined the voice navigation strategy that we will use and conducted an initial usability study on this navigation strategy. From the usability study we have obtained validation for some of our approaches, and learned some new concepts in voice navigation as well. With our findings, we are defining annotation tags that can be used to produce highly usable web pages over a phone user interface. In this paper we describe our initial study and the findings of the study.

1. Introduction

In today's society, continuous access to information has become a necessity. The Internet has become a main source of information, even for traditional information providers such as TV and Radio stations. This information availability in a single medium has driven technology to extend access to internet-based information from many different devices, such as laptop with wireless access, cellular phones with web browsers, PDAs, even some traditional home- appliances have optional internet-connectivity. The next logical extension of the internet is to voice-enable the web so that we can access information from phones and cellular phones that are not equipped with web-browsers. This will extend the reach of web-based information to match the reach of the global phone network. This is a particularly attractive solution since cell phones are very popular and in use in countries and rural areas where connectivity to the internet is difficult to achieve. Furthermore, the mobile phone is already used as a "device proxy" [Perry 2001]. It has become the preferred way to stay in contact with the main office while mobile workers are on travel.

VoiceXML is a markup language created by a consortium of companies that includes Lucent, Motorola, IBM, and AT&T. The language is XML based and requires only a text editor to create VoiceXML pages. The language includes tags that are specifically geared towards building voice interfaces and to support telephony applications. The ease of editing a VoiceXML file and the unique features of the language makes it relatively easy to create a voice interface for phone-based applications. Furthermore, the designers of the language have opted to use the World Wide Web infrastructure for delivery of VoiceXML content. VoiceXML pages are served by an HTTP server (e.g., Apache), and they communicate back to the server via CGIs and URL, much like the web does. Furthermore, VoiceXML supports the use of EMACScript to do specialized processing within a VoiceXML page. The promise of a voice-enabled web is discussed in some detail in two recent articles [Danielsen 2001, Lucas 2000].

How do we take advantage of existing web infrastructure?
Given that browsing the web by phone seems like a very possible avenue for extending the reach of web-based information, it is necessary to investigate how the information contained on web pages should be presented to the user. The information should be presented in a clear, easy to navigate manner. One would also hope that the presentation of information in a voice user interface could be standardized in some way so that the user does not have to learn new interaction and navigation techniques for each web page they visit with a phone. Much like the basic controls in a web browser are sufficient for most navigation tasks, we need to identify their equivalent in a voice browser. Research needs to be done on the best ways to present the sometimes complex structure of a web

page to a user so that they can navigate it easily.

There are two approaches to present web-based information over the phone. The first one, usually called Transcoding, uses software to convert existing HTML content into a form that is suitable for presentation in a phone-based user interface (e.g. VoiceXML). Any web-based information that exists at a website is usually available over the phone in this approach. However, the usability of this type of service is very poor. Websites are currently defined in such a way that they rely heavily on the visual scanning capabilities that humans have. Lots of information in a web page has ordering and structure that is visible to the human eye but not necessarily identified in the HTML code (e.g. frames, HTML tables). As a result, conversion of these pages to VoiceXML is a very difficult problem (if not an impossible one). The result of most of these automatic transcoding algorithms is a generic voice interface that usually has low usability.

A variation of the transcoding approach is to use annotations that aid in the conversion from HTML to VoiceXML. The annotations are extra HTML-like tags that are added to an HTML document so that guide the conversion process. These annotations are defined in such a way that the transcoding of the original document can produce a VoiceXML file with higher usability.

Services that use automatic conversion and transcoding often have the following characteristics:

- convert web context to VoiceXML; often based on annotations/tagging
- usability is often low for automatic conversion and better for annotated systems
- user knows there is an existing structure (e.g. the web space) but might be difficult to map the two
- getting lost in structure is a problem; there is no single "main page" or easily identifiable landmarks
- some of the tags in previous work have been defined without usability goals in mind; thus many tags identify information to remove from the voice interface but do not indicate how to use information in the voice user interface (e.g. is the information part of a prompt or just textual information presented to the user?).

Another approach to serve web-based information over a phone-user interface is to use a centralized service. The personnel at the centralized service write specialized programs that provide information gathered from the web over the phone. This usually requires some form of coding/annotation on their part. This approach has higher usability than the transcoding approach but the information available is restricted to only those web pages that the central service has "converted." Several companies have begun offering services based on this approach, such as BeVocal and TellMe.

Information services such as BeVocal *and* Tell-Me have the following characteristics.

- information services process information from the web and other sources and make it available via their voice system
- user do not know structure of information being browsed
- lost in navigation is not a big issue, since structure is hidden
- user is active requesting information, information retrieval task
- most successful right now
- cost intensive, none or little support for in-house voice publishing

Our Work

This report describes our initial study to evaluate how to navigate web spaces using a phone-based voice user interface. We are exploring the usability aspects related with voice navigation of web spaces. With the findings from this study, we hope to define some annotation tags that can be used to automatically generate highly usable voice user interfaces based on web pages.

For the particular study presented here, we had four goals. We defined a navigation strategy for highly-structured pages, such as the My Yahoo! page [Manber 2000]. These types of pages include several subsections, with a specific type of information in each. Furthermore, there is little or no relationship between one subsection and the next. For this domain, we defined the following evaluation goals:

1. First Impression - training should not be required to use our system. As such, we were interested in the user's initial reaction on how to use the system. Voice interfaces have a reputation to be fragile and error-prone. We wanted to have a system that was easy to use for newcomers.
2. User Satisfaction - we wanted the users of our voice navigation system to feel good about their experience using the system. We wanted the users to want to use the system again if the opportunity was available.
3. Prior Knowledge - since we were using an existing web page we were interested to know the effect that prior use of My Yahoo! would have on voice navigation. We expect that prior experience with My Yahoo! would make voice navigation of the same space easier.
4. Mental Model - we wanted to know what mental model the users developed after using the voice interface. In particular, we wanted to know if users form a mental model of the web space that is similar to the structure of the web pages (textual web pages).

This report describes the study we conducted to explore these four usability goals. Also, towards the end we provide a new design of our voice navigation strategy for this particular domain, based on the lessons learned during this evaluation.

2. Previous Research in Voice User Interfaces for the Web

Voice Navigation

Previous work in the area of voice navigation for the web has focused on browsing web pages by phone. Goose, et al. [Wang 2001] implemented a system that allows the user to store a set number of pages that you wish to visit on a server. When the user calls into the system, the user can choose the page he/she wishes to browse. The information is presented to the user by parsing the document using a text-to-speech synthesizer with features such as multiple voices, pausing, etc. to make the document more interesting and comprehensible. Links are indicated by having a unique sound effect followed by a voice reserved specifically for announcing links. With this system, a user may select a link at any time until the next link is presented. This system also has three other modes in which the user can hear just the links, just the section headings or just the content of the web page, as well as controlling the presentation with "rewind", "fast forward", or "pause."

Poon and Nunn [Poon 2001] designed a system that incorporated voice and keypad presses. The system allows users to browse any web page, not just a set predefined few. To go to a new web page, the users spell out the address of the web page. Poon and Nunn's navigation techniques include capabilities for navigating within a page as well as between pages, having bookmarks and a history list and also following hyperlinks. The system uses text-to-speech to read the page but also includes sounds to indicate document structure. These sounds are ones such as a camera shutter clicking indicating an image, or a doorbell indicating an e-mail address. In this way, the user gains a greater understanding of the structure of the page. They found that while sounds indicating structure where helpful, they were also distracting.

Borges, et. al [Borges 1999] did an interesting study on whether users can effectively use speech to browse a visual web page and whether they preferred that method over the mouse. From their experiments, they noted that users tend to use short one or two word phrases to navigate and prefer interfaces where a small vocabulary is used.

Audio Presentation of Links

Other research has been done to determine how to present a hyperlink in a voice user interface. A study by Wang, et al. [Wang 2001] examined the effects of speaker change, volume change, link position and link length on the detectability and comprehensibility of a link within text. They found that it is much easier to detect and understand a hyperlink when it is presented in a pre-recorded human voice whether it is indicated by a change in voice or a change in volume. When a pre-recorded voice was used, a change in speaker was better for indicating a hyperlink than changing the volume. However, for a text-to-speech system, links are better detected with a speaker change. As for link position, overall, a link is more easily detected when in the middle of a sentence. Particularly, a link is more easily understood in the middle when volume change is used but is not easily understood when speaker change markup is used. Also, they found that longer links improve detectability when using speaker change with text- to-speech but also provided lower comprehensibility.

Voice User Interface Design Guidelines

Yankelovich [Yankelovich 1996] presents different types of prompt designs that should be used with voice system. She discussed the difficulty in designing the prompting for a system since the users will often assume they can use phrases that the system does not support and at the same time they do not realize all the phrases that the system

does support. The difficulty arises in determining how much guidance the prompts should give the user. According to Matt Marx of AITech, there is a continuum from implicit to explicit prompts upon which different prompts fall. Explicit prompts are best for systems with a small vocabulary. An example is a directive prompt which tells the user the exact words they should say [Yankelovich 1996, p. 37].

Implicit prompts are better for systems with large grammars and many options. With these types of systems, users can use conversation-like interaction with the system. Users can speak sentences with these systems instead of just phrases and the system may ask them for any information that they do not automatically provide. Systems can also prompt users on what to say implicitly by the language used in the prompts since people tend to imitate the speech patterns of the person to whom they are speaking.

Two techniques that fall in the middle of the spectrum are incremental and expanded prompts. With incremental prompts, the user is provided with an implicit prompt. The system will wait for a response and will provide more explicit information if the user does not give the desired response or if the user remains silent. Expanded prompts are similar to incremental prompts with the difference that they require that the user respond in some way to the initial prompt and if this response is not interpretable, the user is then presented with an expanded version of the first prompt. Both these techniques work well for systems that have frequent users, since they will not be slowed down by unnecessary prompts.

Another technique that falls in the middle of the spectrum is that of tapering. The goal of using tapering is to shorten the time of interaction with the system as users gain more experience using the system (within the same session). The first time a user encounters a section of the system, they are presented with an explicit prompt. If they access that same section again during the same session, the prompt is less explicit. Another method designed to reduce interaction time is that of using hints. With this technique, the user is first presented with prompts that provide an "implicit conversational question followed by an explicit hint" [Yankelovich 1996, p. 41]. Hints only include a few of the possible choices and are designed to be removed as the user gains more experience with the system.

The research discussed here clearly shows that there is behavior in a voice user interface that must be supported but that is not part of the original HTML file available on the web. This is the single most important reason why simple syntactical transformations are not enough to produce a good voice user interface. Some of the issues in a voice user interface that needs to be addressed in the tags or transformation algorithms include:

- how to present links
- how to convey page/information structure
- what voice commands to use for navigation within a page and across page boundaries
- interface style to use, such as implicit/explicit prompts, tapering, etc.

3. Voice Navigation of My Yahoo!

We selected My Yahoo! because it is a typical site of many other personalizable portals on the web. They all have independent sections, each clearly delineated visually on the screen. The visual structure of this type of pages is a perfect candidate for voice menu presentation to an user. We also wanted a site that some people had seen and we thought that My Yahoo! was a good candidate for that.

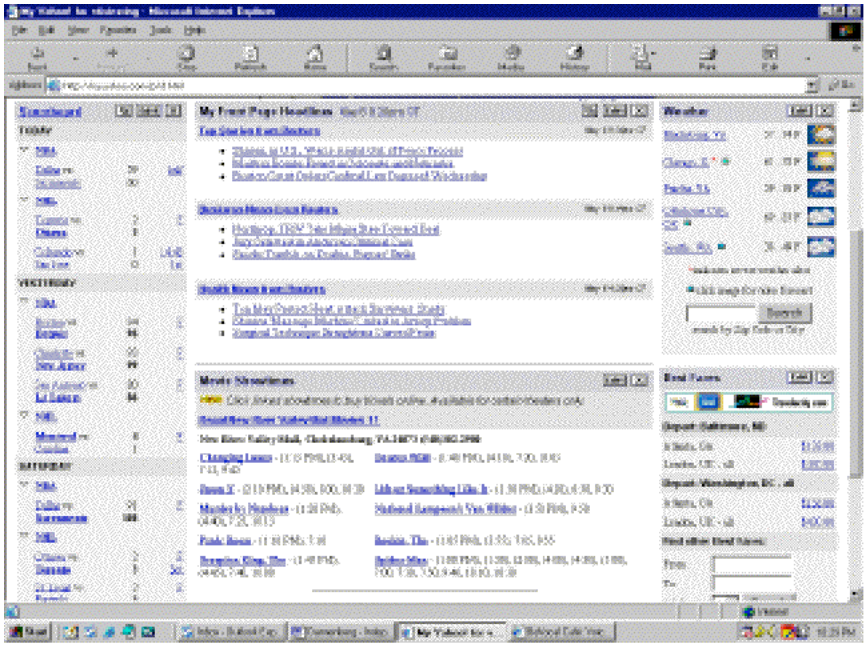
We divided the participants in three groups, those that had never seen My Yahoo!, those that had seen the page before but never customized it, and those that have customized My Yahoo!. Based on Manber [Manber 2000], most users of My Yahoo! never customize their site, so the middle group is a significant group to consider.

Information contained on pages such as My Yahoo! has several distinct characteristics important to browsing it via the phone. First of all, My Yahoo! is customizable so that the user can determine which modules of information they would like to see when they log into the web page. In this way, the user sees only the information in which they are interested. For instance, the user can select which news modules they would like to be displayed on the page. They can choose from categories such as "News and Politics", "Business and Industry", "Commentary", "Entertainment", "Health", "Music", "Sports", "News from Europe", etc. They can also choose to receive daily articles on topics of their choice written by experts in that field. They can set up links to shopping or chatting, and even search engines. They can also have the page alert them when they have new mail on Yahoo! One of the attractions to such a page is that users can configure it to contain most of the information or links to the information to which they need access on a regular basis. In this way, users can use this page daily if not hourly for easy access to such information. Another feature of this site that is significant to browsing the page by phone is that

users can customize to some degree the way the information is presented to them visually. In this way, they can put the more important or more frequently used information towards the top of the display, or wherever they decide it is most relevant or easily accessed.

Another feature of My Yahoo! is that information is grouped in general modules such as “Headlines”, “Weather”, “Movie Show Times”. Each of these modules then contains further options that group the information even more. Under “Headlines” the user can choose “Top Stories from AP”, “Science New from AP”, or “Health News from AP” (all of those choices are customizable). Then, under each of these categories, the user can select a specific story. So, in essence, the information is presented in visual groups on the page and there are groups within groups. This is something that is easy to translate into a hierarchical voice menu.

FIGURE 1. My Yahoo! Web Page



All of these features are helpful given that normally, when we view a web page, we can quickly glance over all the sections contained on the page to get an overview of everything that is there. Over the phone, this is not as easily accomplished. If the page is unfamiliar, we will have to listen to a relatively large amount of information to get an idea of what information is contained on the page. So, My Yahoo! has the advantage of allowing the user to set up visually the information that will be presented to them in the order that best suites them as well as control the information contained within the page so as to limit the information that is presented to them, thus also saving the user time by not having to listen to information in which they are not interested.

Navigation Strategy: Initial design

We decided that the best way to present information to the user would be in a hierarchical levels. We felt that this matched well with the visual grouping method used by My Yahoo! Since My Yahoo! is already split up into sections (see Figure 1), we used the titles of each of these sections as menu items. We decided that the user should first be presented with the name of each module such as "Headline", "Weather", or "Movie Show Times" in the main menu. We decided that these modules should be ordered, for the most part, in the menu based on how they were presented visually on the My Yahoo! web page if read in a left to right, top to bottom fashion (based on the default My Yahoo! page). When one of these main links is chosen, a new menu is then presented listing the choices in that group as read top to bottom as presented in the visual web page. The process was repeated until the user obtained the information for which they were looking. In essence, when the user follows a link, they are either taken to a new menu and given more options for further refining what they would like to hear (in the case of hearing the weather for today or the 5 day forecast) or are immediately read the related information. Thus, the user has to "drill down" through at least two levels, sometimes as many as four or five levels, before they obtain the information they seek

TABLE 1. Menu Structure used in the Phone-based Voice User Interface for this evaluation

Level 1	Level 2	Level 3	Level 4
Main Menu	Headlines	Top Stories	U. S. tries to stop Bin Laden escape
			Missing, dead at World Trade Center drops below 3900
			Navy to stop ships off Pakistan coast
		Business	Stocks slip further from bull market
			After the bell techs slip eyes on Meade
			Jobless claims fall, consumers upbeat
		Science	FDA approves first sepsis drug
			Researchers prone first space sleep
			Study: thieving birds watch their back
		Health	FDA approves first sepsis drug
			Connecticut woman dies from inhaled anthrax
			Flu shot found safe for asthmatics
	Scoreboard	Today's Scores	NHL | NBA
		Yesterday's Scores	NHL | NBA
	Weather	Blacksburg	detailed | extended
		Fairfax	detailed | extended
		Seattle	detailed | extended
		Chicago	detailed | extended
		Oklahoma City	detailed | extended
	Movies	(this was a passive choice, users just heard the movie listings)	
	Best Fares	Departure City: Baltimore	Arrival City: London | Arrival City: Atlanta
		Departure City: Washington D.C.	Arrival City: London | Arrival City: Atlanta

.

An example of this is found in the following sample transcript:

Computer: Welcome to My Portal! At any time while using this system, you may say main menu. Please say one of headlines, scoreboard, weather, movies or best fares.

User: Headlines

Computer: Headlines! Please select from the following headline items: Top Stories, Business, Science, Health

User: Science

Computer: Science. Please select one of the following science stories: FDA approves first sepsis drug, Researchers prone first space sleep, Study: thieving birds watch their back

User: <Silence>

Computer: Please say one of: FDA approves first sepsis drug, Researchers prone first space sleep, Study: thieving birds watch their back

User: FDA approves

Computer: <reads article> To go back to the main menu, say main menu. To go back to health, say health. To go back to science, say science.

4. Method

Using VoiceXML, we implemented a mock up of a simplified My Yahoo! web page. The sections contained under the "Main Menu" are: Headlines, Scoreboard, Weather, Movies and Best Fares. We chose these based on the fact that we were mainly focused on the navigation technique and did not want to overwhelm the user with choices but also wanted to present topics that a wide variety of people would most likely be interested in and thus have on their own My Yahoo! page. Once the user selects one of these topics, they are taken to another menu and presented with another list of selections and so on until they reach the information they are seeking.

We used explicit prompts for all of the menus except the weather where we used incremental prompts. Incremental prompts were used in particular for the weather since the cities for which one checks the weather is fairly constant and the user is likely to remember which cities they have selected on their My Yahoo! web page.

Usability goals

As mentioned before, the usability goals for the evaluation were:

- user satisfaction
- initial reaction
- what effect does prior use of My Yahoo would have on voice navigation
- what type of mental model of the site the user developed after using the voice interface

The organization of the menus is shown in Table 1. The user has to navigate following the hierarchical levels. For instance, to get to today's scores for the NHL, she first has to choose "Scoreboard" from the "Main menu". Then she has to choose "Today's Scores" from the "Scoreboard" menu and finally, "NHL" from the "Today's Scores" menu. To navigate backwards, the user can either say the name of the previous level or say "Main menu" to be taken back to the main menu. She is unable to just say the name of a level anywhere in the system and be taken there unless that level is either one higher or one lower than her current level. Not shown in the table above is a final level that contains the actual information the user is seeking. For instance, there is a level 5 for the Best Fares menus that contains the plane fare from Baltimore to London.

It is interesting to note that this method did not map directly to the links contained in My Yahoo!. For instance, much of this information above was contained within the first page of My Yahoo! Information such as Scoreboards, Movies and Best Fares could be seen at a glance on My Yahoo! as this information was contained on the main page. However, information such as actually reading the headline stories or seeing the weather for a city involves clicking on a link from the title of the story, which will then take the user to another page where the story is contained.

5. Usability Evaluation

A total of eleven (11) participants used the system. All users were Virginia Tech students and were recruited from several Computer Science classes. The system was built with VoiceXML running on BeVocal's system (www.bevocal.com). The users were given four tasks to complete and then completed a survey evaluating their experience with the system.

Table 2 shows one set of the tasks used. There were several sets, the difference between them was all in the details of the information being sought. For example, task 4 was always finding the results of a basketball or hockey game, but the teams changed from one task set to the next

TABLE 2. Tasks done in the experiment

Task description
1. Listen to today's weather forecast for Blacksburg, VA. What will today's high be in Blacksburg? What is the current temperature in Blacksburg?
2. What is the plane fare from D.C. to Atlanta?
3. Who won today's basketball game between Detroit and Philadelphia?
4. Listen to one headline that interests you. What is the article about?

Three of the four tasks were very specific, allowing the user little room for exploration. An example of this is "What is the fare from Washington D.C. to Atlanta?". The fourth task was a less specific in that the users were asked to find one news article in which they were interested and listen to it. This was designed to allow them to have a little bit of flexibility in terms of where they went in the system and yet was still structured enough to be realistic of a task they might perform in actual every day usage.

Users arrived to our lab and sat in front of a touch-tone desk phone. They dialed the BeVocal phone number and entered an ID and a PIN supplied by the experimenter. On average, the time the users spent connected to the system was around 7 minutes. We realize that users might not actually spend that much time on a system like this, but we wanted to allow the users enough time to navigate through the system and to perform all the tasks required for the usability evaluation.

The experimenter recorded any errors or critical incidents that occurred while the participants were using the system. After completing the tasks, the users then completed a questionnaire regarding their experience with the system and the experimenter allowed them to express any other comments about the system.

6. Results

Of the eleven (11) participants, four (4) had never used My Yahoo! before, three (3) had used it but not customized it, and four (4) had used and customized it. Table shows a summary of the results. We did not have enough participants to perform formal statistical analysis (this was not the purpose of this evaluation), instead we show descriptive statistics based on the user's responses to the questionnaire and observations made by the experimenter. Table 3 shows the results from the questionnaire for all three groups.

TABLE 3. Summary of questionnaire results.

Numbers are average responses to questions on a scale from 0 to 10	Never Used (4 partic.)	Used but not customized (3 partic.)	Used and Customized (4 partic.)	Overall Average (11 partic.)	Figure
Navigating the system: Not comfortable/very comfortable	8.25 (+)	6.3 (0)	7.75 (+)	7.5 (+)	See Figure 2
Navigating the system: Very easy/ very difficult	0.5 (+)	0.6 (+)	3 (+)	1.4 (+)	See Figure 3
Menus: Not helpful/very helpful	8 (+)	6.3 (0)	8.5 (+)	7.7 (+)	not shown
Accomplish tasks: Very easy/ very difficult	4.75 (0)	3.6 (0)	4.25 (0)	4.2 (0)	See Figure 4
Locate sections: Very easy/ very difficult	5 (0)	7.3 (+)	4.25 (0)	4.3 (0)	See Figure 5
Correct mistakes: Very easy/ very difficult	1.25 (+)	6.66 (0)	4.66 (0)	3.9 (0)	See Figure 6
Reaction to system					
Frustrated/ satisfied	6.5 (0)	4 (0)	7.5 (+)	6.2 (0)	See Figure 7
Very easy/ very difficult	2 (+)	3.6 (0)	2 (+)	2.5 (+)	See Figure 8
Simple/Complex	1.25 (+)	2.3 (+)	3 (+)	2.2 (+)	See Figure 9
Satisfaction with system: Not satisfied/ very satisfied	7.75 (+)	5.3 (0)	7.25 (+)	7.1 (+)	See Figure 10

The numbers show are averages for each of the groups. The table cells show an indicator of whether the average represents a favorable (+) opinion, a neutral (0), or a negative opinion (-). These were determined just for classification

purposes based on the following scales. For neutral, we used score > 3.0 and score < 7.0. The positive and negative were on the two sides, depending on whether the scale was positive (that is higher numbers meat a positive response, e.g. see first question) or negative (higher number meant a negative response, e.g. see the fifth question).

Below we have included graphs showing these numbers in bar graphs for each of the questions.

FIGURE 2. How comfortable were you navigating through the system? Not comfortable (0)... Very Comfortable (10)

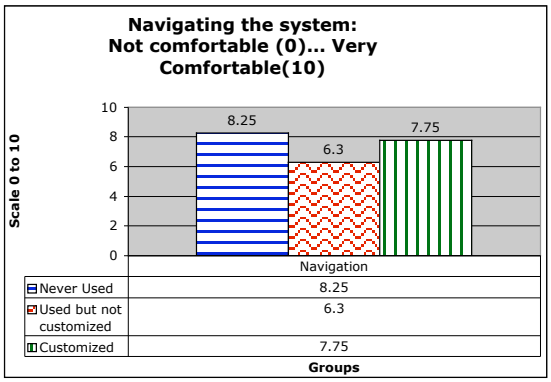


FIGURE 3. How difficult was it to learn how to navigate through the system? Very easy (0)... Very difficult (10)

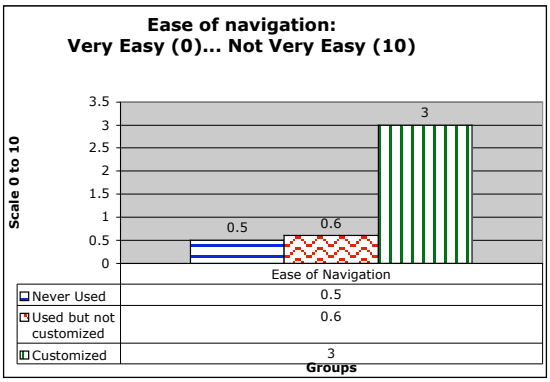


FIGURE 4. How easy was it for you to accomplish the tasks given? Very easy (0)... Very difficult (10)

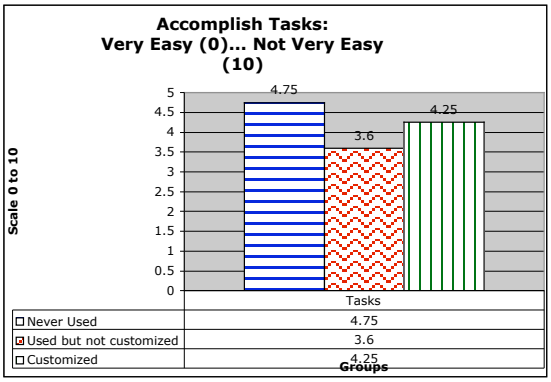


FIGURE 5. How easy was it for you to find the sections required? Very easy (0)... Very difficult (10)

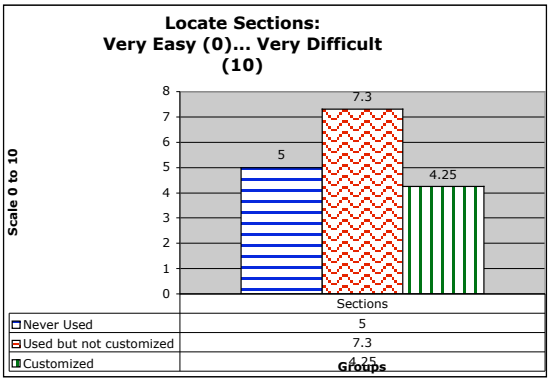


FIGURE 6. How difficult was it to correct a mistake you'd made? Very easy (0)... Very difficult (10)

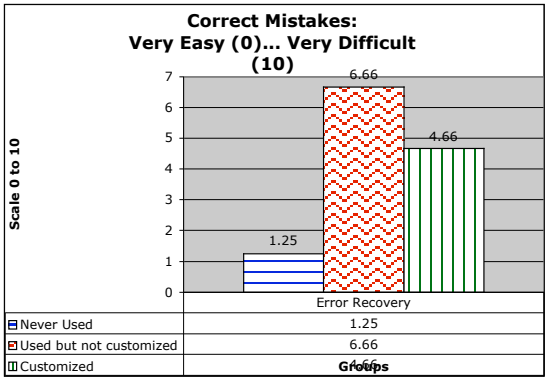


FIGURE 7. Rate your overall reactions to the system in regards to: Frustration vs. satisfaction. Frustrated (0)... Satisfied (10)

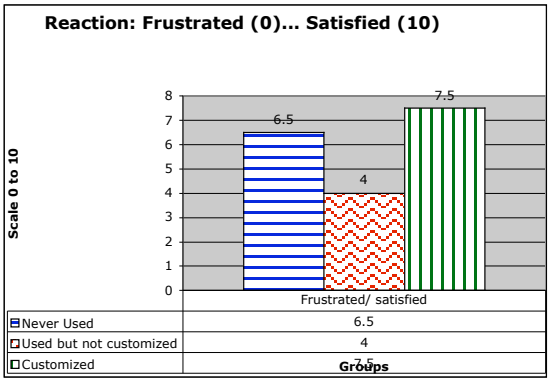


FIGURE 8. Rate your overall reactions to the system in regards to: Difficult vs. easy. Very easy (0)... Very difficult (10)

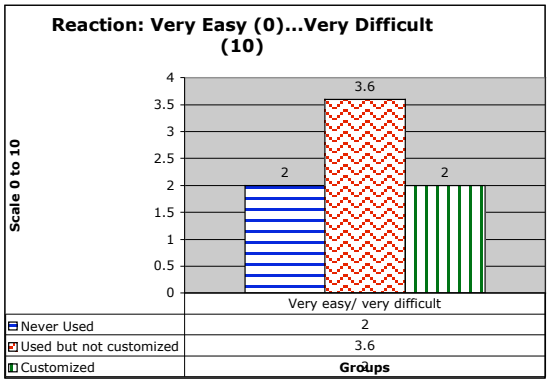


FIGURE 9. Rate your overall reactions to the system in regards to: Simple vs. complex. Simple (0)... Complex (10)

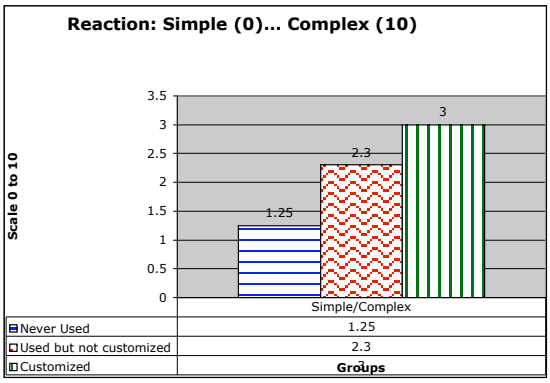
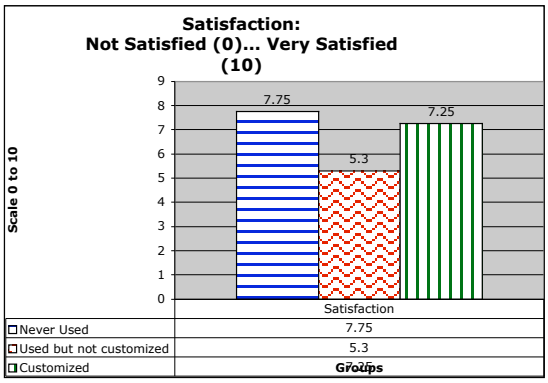


FIGURE 10. How satisfied were you with the system? Not satisfied (0)... Very satisfied (10)



7. Discussion

We used the usability evaluation to answer the following four usability questions (goals):

- what was their initial reaction after using the system?
- how satisfied were the users of this system?
- what effect did prior use of My Yahoo have on voice navigation of the same space?
- what type of mental model of the site the user developed after using the voice interface?

Based on the positive/neutral/negative classification done, we found that all responses to the questionnaire were either neutral or positive. Four questions produced neutral results and six produced positive responses. There were no negative responses (on average). This is a good indication that the system was easy to use and well received. However, there a number of suggestions and findings that will allow us to make the system easier to use. This section discusses all of these results. As a word of caution, we did not have enough participants to make any statistically valid claims, so all comments here are made as trends and early evaluations of the voice interface here described.

User Satisfaction

In regards to the overall system, on the average users found it more satisfying than frustrating (avg=6.2, see Figure 7), but according to our classification, we deem this to be "neutral." Participants who had customize My Yahoo!

were more satisfied (avg=7.5) than participants in the other two groups.

On average, participants found the system simple and particularly easy to use (avg=2.2, Figure 9). Participants that had customized My Yahoo! before, found it slightly more complex (avg=3.0) than those who had not used My Yahoo! or had used but not customized My Yahoo!

On average, participants were satisfied with the system (avg=7.1, Figure 10). It is interesting to note, that once again those participants that had never customized My Yahoo! before expressed almost neutral opinion with regards to their satisfaction (avg=5.3).

There were several comments that had an effect on user's satisfaction. Many of them had to do with navigation of the menu system, discussed below. Also one person was concerned with whether or not they would be able to understand the voice in a crowded area. There were several comments regarding what we classify as "technology satisfaction", that is user satisfaction as a reaction to the particular technology used (e.g. poor voice synthesis, not recognizing barge-in, etc.). These comments are discussed below in a separate section.

Finally, we asked all participants an open ended question: Would you use a system like this? Typical responses were:

- "No, all these things can be done better at other times"
- "Only if other media were unavailable and I had a phone".
- Several said they would use it to check the weather, or movie listings or sports scores.
- Another said he would use it if it was free.

Initial Reaction

While overall users had a positive reaction to the system, the questions that received neutral responses all had to do with the use of the system. Remember, that they had received no previous training with the system and received no particular instructions on how to navigate the menu system. Overall, most people were comfortable navigating through the system with those who had never used My Yahoo! tending to be slightly less comfortable.

All users were neutral on whether it was easy to accomplish their tasks (avg=4.2, Figure 4). They were neutral also on how easy it was to locate the different sections of the My Yahoo! space (avg=4.3, Figure 5). It interesting to note that those participants that had never customized My Yahoo! classified locating sections as a problem area (avg=7.3 where 10 was very difficult). This was the only negative result in this evaluation.

Also, on average all users were neutral on the difficulty to correct mistakes (avg=3.9, see Figure 6). Only the group that had never used My Yahoo! found it easy to correct mistakes (avg=1.25).

Prior Experience with My Yahoo!

Prior experience with My Yahoo! seems to be a determining factor for the results we collected here. Notice, that we do not have sufficient data points to make statistically significant claims, but there seem to be some common trends verified by several questions and even by one question posed twice, one as a positive and one as a negative question. With this in mind, here are the results.

Users who had never used My Yahoo! before, rated 7 of the 10 questions as positive (+ shown on Table 3). Users who had used My Yahoo! but not customized it had only 2 positive responses, 1 negative, and seven neutral. Note that this comprises the larger group of My Yahoo! users on the web [Manber 2000]. The last group, those that had used My Yahoo! and customized it rated 7 questions as positive and 3 as neutral, just like the group that had never seen My Yahoo!

Our results found that those that had used My Yahoo! (but not customized it) and those that had never used My Yahoo! found the system easier to use. For one group, those that never seen My Yahoo! before, the system was just another voice over phone user interface. Since they had no prior experience with My Yahoo!, no interference was expected. For those that have used My Yahoo! and never customized it, the system seems like a good voice user interface to the information available in the default My Yahoo! page.

However, for those users that knew the most of My Yahoo!, that is those that had used and customized the content of their My Yahoo! pages, evaluated our voice user interface with either neutral or negative opinions. So it seems, with our brief evaluation, that we have identified an easy to use voice user interface, but one that does not scale well for more "domain knowledgeable users." This will be discussed in more detailed below.

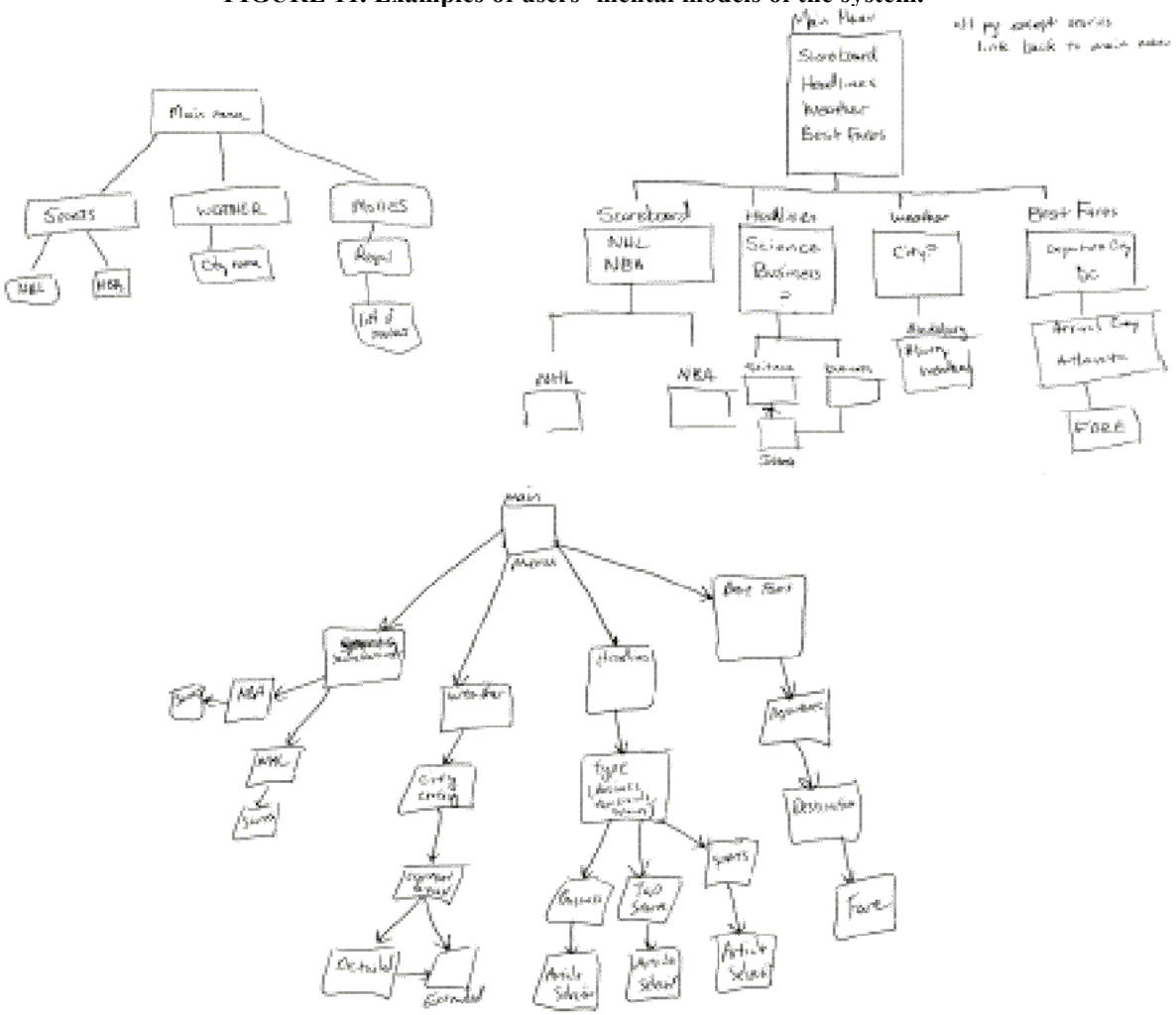
Mental Model of Site

Browsing the web using a traditional desktop browser produces certain "mental model" of how web sites are organized and how the web as a whole is organized. We often use this model to re-find information and/or to navigate previously visited sites.

We wanted to explore what mental model users would develop from having used the site. To that end, we asked all participants to draw a map of the structure of the site.

It was very interesting to find that all users thought of the system in terms of a series of links from one page to the next. None thought the "voice site" was related to a single page on the web, even though the basic My Yahoo! page is a single portal page. None of them related the system to what an actual My Yahoo! or other web page might look like in terms of having multiple sections on each page. In particular, the information that we used in the voice interface is information that comes from the basic default My Yahoo! page.

FIGURE 11. Examples of users' mental models of the system.



All participants drew a hierarchical website structure that matched the menu structure they used in the voice site. Figure 11 shows three examples of what users drew as their understanding of the website structure. This finding has important implications for the design of voice interfaces to already existing application domains. In the particular case of the WWW, users see the voice client and the traditional client as clients to different applications, even if the underlying data is the same. They failed to make a connection of the structure from one site to the next. In the voice interfaces, the structure is explicitly manifested as the voice menu. In the traditional browser, however, the structure is implicit in the way the site is laid out. As can be seen on Figure 1, the graphical interface uses color to highlight the title of each section. Humans highly developed visual scanning abilities allows us to see this structure. However

that structure does not seem to be perceived as a structure for navigation in the voice interface, even for those participants that had seen My Yahoo! before.

Navigation, Use of Menu

Overall, most people were comfortable navigating through the system, as indicated by their answers to question 1 (avg=7.5, see Table 3). While menus were considered helpful (avg=7.7), locating the different sections of the voice site was neutral (avg=4.3), and correcting mistakes was also neutral (avg=3.9). So, clearly there is room for improvement on how navigation was handled in this initial voice interface. This section, based mostly on user comments, discusses the findings regarding navigation and use of the menus.

Back

Two participants wished the system had the ability to go back one step in the system by saying "Back" no matter what your location. The system had a "main menu" command that would take the user back to the top level menu (see Table 1), but no "back" command was implemented. This was complicated with the fact that some users did not like having to go back to the main menu all the time, several complained that the greeting was always the same.

Break

Voice user interfaces need to make effective use of "barge-in." Unfortunately, for our usability evaluation, the barge-in functionality was not working. The result was that, as expected, users found it frustrating not being able to break out of "long" stories. Many participants found the navigation of news stories confusing because they were not able to break out like they expected. Also, while using the menus, they had to wait until the system was done talking before they could issue commands.

We suspected that this would be a problem. To alleviate its effect, the news articles were cut in length by at least half and sometimes by as much as by three-fourths in order to make them relatively short. Even with the reduction in length, all users became impatient waiting for the article to finish. All wanted to be able to break out of the system and go back to the main menu. While part of this was a flaw with BeVocal's implementation of VoiceXML (or our understanding of VoiceXML), it does bring up a good observation. When using the phone to access information, users appear to prefer short summaries of information and do not want listen to long excerpts.

Users did not like hearing main prompt over and over again: "Welcome to my portal. At any time while using this system, you may say main menu." This could have been alleviated by implementing tapering prompting techniques as mentioned by Yankelovich [Yankelovich 1996].

Forward/sideways jumps

When navigating through the system, some users tried to side-step the menus by speaking the choice they wanted before they actually heard it. A few times, when the users did this, they used words commonly used in everyday speech to describe what they wanted instead of using words that My Yahoo! contains. For instance, some users said "hockey" instead of NHL and "sports" instead of "scoreboard." This indicates that perhaps menu items should have names commonly used in everyday speech to facilitate recollection. At least, the grammar should allow these commonly used words to refer to the proper page section. Another similar incident occurred when one user tried to say "menu" instead of "main menu" to go back to the main several times in a row before figuring out that they had to say "main menu". Perhaps more studies could be done to find out if over time, users remember the terms used by My Yahoo! and adapt accordingly when using the phone system or if they will always think in terms of everyday words for such things.

One participant wanted direct access to the information without having to use menus. In essence, s/he wanted to be able to tell the system what s/he wanted without having to go through all the menus. This somewhat corresponds to the hints technique of prompting described by Yankelovich (1996).

Others wanted to be able to move easily between selections under a menu. For instance, instead of having to go back to the main menu of sports, users wanted to be able to navigate to the NHL scores from the NBA scores. This represent "sideway" navigation of menu structures.

One user wanted to avoid problems associated with voice menu selection by trying to use the keypad. This particular user wanted the ability to press buttons for choices instead of having to speak the choices. This could be easily added to our prototype, since VoiceXML does support the use of the phone keypad.

Another person thought that the way the scores were arranged was confusing. Instead of choosing "Scoreboard" then "Today's Scores" or "Yesterday's Scores" and then the sport they wanted to hear, the would have preferred to have been able to choose "Scoreboard" then "NBA" or "NHL" and then hear either today or yesterday's scores.

Some users also had problems in the weather section. Some users did not know how to find the highs and lows for a city. Instead of saying "detailed" or "extended" like they were prompted to do, they said the name of the city again. It took several tries before one user figured out what s/he needed to say.

Menu labeling

Some users were confused as to how to select items with more than one or two words such as in headlines. For instance, one of the headlines was: "Navy to stop ships off Pakistan coast". A few users were unsure as to whether they had to say the whole headline or only part of it. This confusion probably came from the fact that users tend to mimic the speech patterns of those to whom they are listening [Yankelovich 1996]. However, when the system spoke the entire title of the headline, they were unable to remember all of it word for word and thus got confused as to what they should say. Most just said one or two words even if they were unsure. This also supports the work by Borges, et al. [Borges 1999], in that the users tended to use short one or two word phrases to navigate.

Two users tried to say the name of a movie to hear the times for it instead of having to listen to all the movies. We found this particularly interesting given that the way one normally finds out information for movies via the phone is by calling the movie theater and listening to a recording of all the movies and their times. However, this could have been due to the users' previous experience interacting with the weather module that allows them to just say the name of the city for which they want the weather. Given that users expected to be able to say the name of a movie indicates that they must have had some concept of the information being gathered (movie listings) and thus thought that they could access that information directly without waiting for a long information playback. Other users that did not try to say the name of the movie commented that they wanted this ability.

One interesting finding that we learned, which will be discussed in a later section, is that there are different usability expectations for different "types" of menus. We have identified 3 types of menus, each with slightly different usability requirements for voice navigation. We will use these findings to define our annotation tags that will help generate VoiceXML automatically from HTML. Some of the differences are based on whether the user knows the contents of the menu ahead of time (e.g. the Sports section) or not (e.g. the news stories). If the user knows ahead of time, our participants wanted to be able to say that section out of turn and to be able to jump to that menu item directly. Also, if they knew the structure of a higher menu than the one they were listening, they wanted to do a sideways navigation (e.g. moving from NBA to NHL scores). Finally, if the menu contained "structured" information (e.g. the movie listings), they wanted to be able to "filter" the information. For example, in the case of the movie listings, they wanted to be able to hear information about a particular the movie, not all the movie listings.

Technology problems

We found a small number of problems that were all related to the particular VoiceXML technology use for the evaluation. We expect this technology to get better with time and thus some or most of these usability issues should disappear. Nevertheless, given the state of the art of the technology now, we feel that these are issues that will slow down the adoption of this new voice technology.

Seven participants found that it was difficult for them to hear and understand the menu choices. They found the text- to-speech voice difficult to understand at times due to both the pronunciation and the speaking rate, thus making it hard to navigate the system. Some users expressed interest in wanting to be able to control the rate of speech of the system.

Also, as mentioned earlier, the barge-in feature of VoiceXML was not working. It is not clear if this was a problem with BeVocal's implementation of VoiceXML, or a problem with our implementation of this application in VoiceXML. But, clearly this was a problem for users, and it was one of the most frequently requested features.

Two other comments we found interesting. One user wanted to be able to place a phone call with the system or get 411 type information. But more interesting was the fact that several users did not know how to hang up. One user went back to the main menu and listened to it all over again to see if there were instructions on how to end the call. Other users turned to the experimenter and asked, "Can I hang up on it?". This shows some misunderstanding of the technology.

8. Further Observations

Overall, participants seemed to be satisfied with the system. They found navigation fairly easy with just a few exceptions such as not being able to break out of long articles and not being able to back up just one level at any time. Of those features that the system was lacking, one of the most important ones was that of a customizable text-to-speech speed. Some users thought the system spoke too slow, while others thought it spoke too fast.

In terms of navigation, some of the most important things learned were that users desire quick navigation. They do not like to listen to long lists of choices. If they know the section they want, they prefer to just say the name of that section and have the system take them there. This corresponds with Yankelovich's Hints prompting technique. However, our concern is that if this ability is provided, users may find that they can easily get lost in the system if they are not well acquainted with its structure. Also, users are concerned with retrieving information quickly and in an abbreviated manner. Users do not want to have to listen to a long list of movies when they only want to know the times for one movie and users do not want to listen to an entire news article. They would rather just hear a summary of the article. Overall, users seem to want to be able to use a system such as this to quickly get access to information without having to spend a lot of time on the phone. Part of the system could be improved by decreasing the amount of time spent on prompts by implementing a tapering or incremental prompting scheme.

It is also important that a system such as this be able to recognize and respond to common words and phrases used to describe things since users seem to expect this. Users wanted to say "Hockey" instead of NHL, for example.

Based on our findings, we would make the following guidelines for building this type of interface.

- If possible, allow the user to have some control over the rate of the text-to-speech parser.
- Barge-in capabilities should be allowed at all places and users should be able to go back one step in the system at any point in time.
- Prompts should be more towards the implicit end of the spectrum with either a tapering or hint technique being used to cut down the amount of time the user spends listening to prompts, especially as they become expert users with the system. Given that users will possibly use this system on a daily basis, a hint technique (Yankelovich) where the hints disappear after the user has mastered the system would be most desirable.
- Users should be able to select all menu items by simply saying one or two words of the menu choice.
- Provide support for use of the keypad, as this could alleviate recognition problems, particularly in those times when there is too much noise in the environment and/or the user is getting many misrecognitions.
- Most importantly, all items that can be "thought" of as a list should be placed on a menu. For instance, the movies could have been placed in a menu from which the user could select the one they wanted to hear, as could the different sports teams. Instead, of having the user listen to several hierarchical menu structures, users should have as much control as possible over what information they hear and have a quick way to get to it.
- Along the same line, when presenting news stories to the user, a way should be found to summarize the information to the user so that as few sentences as possible are used to convey the information.

9. Future Work: Improved Voice Navigation of Web Spaces

We started this work by creating a voice navigation that could be used within a transcoding architecture to convert HTML page to VoiceXML pages. Our goal was to define a voice navigation strategy that would produce a highly usable voice interface. In the process we have identified a number of issues that have little to do with HTML and more to do with usability of voice navigation of menu systems. This describes these findings and our approach to how to use these in our next iteration of this work.

Different Types of Menus

While we knew ahead of time there different types of information presented on a page such as My Yahoo!, we did not expect that these differences could have such a different implication for the voice navigation of these pages. There are menus that are always fixed and thus, the user knows their structure and the "commands or menu items" contained therein. In those cases, the user expected to be able to access those sections directly, independent of where in the menu structure s/he was. Two examples of these are the weather and the sports scores menus.

The sports scores menu, for example, is a menu that does not change much from day to day. Everyday there are

scores from college sports and from the professional (NBA, NFL, MLB, NHL) sport leagues. Within those two submenus, the user knows the names of the teams s/he wants to access, and often these are fixed too. The user comes to check the scores for the Yankees, for example, every day. So, the structure to reach this menu is totally irrelevant.

This menu is typical of services such as TellMe and BeVocal, where at anytime, the user can say "Weather" and will be taken to that section. The result of having this type of menu on your system is that the hierarchical structure of the menu disappears. To the user, the system looks like a very flat menu structure.

There is a variation of this type of menu, one which contains highly structured information but with changes that are a bit more frequently that the sports scores or weather. An example of such menu in our system is that of the movies listing. The user might know or not which movies are showing at their local movie theaters. But they might have an idea of what movies are out or should be out at the time of the call. The results is that the user does want to navigate the menu, but not to hear all the details until s/he picks the one movie of interest.

The third type of menu we encountered is one that is completely dynamic from day to day or even hour to hour, such as the news. There is very little likelihood that a user will come to the system knowing what news story s/he wanted to hear. If the news stories are sub classified by area, then the user will know at least the subcategories, such as Financial News, World News, Nation, etc. The implication of this type of menu is that the user has to hear the news stories headings before deciding which to choose. This requires a hierarchical menu.

So, from our initial exploration, we have concluded that while we could build a strictly hierarchical menu system to explore pages such as My Yahoo!, users will hierarchically navigate some of these but for others they will expect a flat menu system with all options available at the same level. There were even indications that the menu system they wanted is instead a fully connected graph, because many participants expected to traverse sideways on the menu system.

Dynamic Menus

The dynamic menus, such as the news stories, described in the previous section, present another challenge for voice navigation. We must consider that voice interfaces built with VoiceXML have a grammar that must describe the user's voice commands. For the dynamic news stories, it is not clear what should be included in that grammar. In our evaluation, we had the full news story headline, but users had a very difficult time remembering the full headline and repeating it verbatim. Seems like it might be better to have an alternative way to select the stories that do not involve repeating the headline. For example, the system could add a number to each story, and present the menu to the user as:

System: Here are today's Financial headlines
1. WorldCom declares bankruptcy
2. AOL also with accounting troubles
3. Accounting students up in arms about recent crisis

The user could then say "story 1" or "first story." The advantage of this method is that it does not place a heavy burden on users' short term memory, which the current approach does. Another advantage is that you can use the same approach for all news stories. We are currently designing another usability evaluation to evaluate this one approach as well as several other techniques for selecting menu items from dynamic menus.

Browsing

The dynamic aspects of navigation produced by the "flattening" of the menu structure presents some challenges for browsing the menu system. In particular, the "back" command, which so many users wanted in our evaluation, now is ill-defined. Upon issue of a "back" command, does the system return to the previously menu played for the user or to the hierarchically-linked menu (up one level)?

Our intuition says that we should return to the menu previously played for the user, independently of where this menu is on the system. This reinforces the idea that the menu system is a fully connected graph, and the back is just traversing in reverse the path followed to get to the current node.

However, users do know some of the structure of the site, evidence of this is the fact that they wanted to traverse sideways to hear other sport scores, etc. So, maybe we need to support a "back" as well as an "up" command. The usability of these two commands needs to be defined and evaluated.

Better use of Technology

One final area where we learned a lot was that we need to make better use of the VoiceXML technology available. For example, we need to implement many of the ideas presented by Nicole Yankelovich [Yankelovich 1996] regarding hints, tapering, etc. These, unfortunately, do not map directly to VoiceXML tags. Instead, there are approaches how to implement them, but our inexperience with this platform prevented us from building the best interface possible. The same can be said about supporting the keypad for selection of menu items, allowing the user to control the speed of presentation by having "faster" and "slower" voice commands, and using "barge-in" to allow escape out of long stories.

10. Acknowledgments

The work described here was made possible by research grant from IBM. Rob Capra contributed to the initial ideas that defined the usability study described here. Natasha Dannenberg designed and conducted the usability study described here as part of her Master's degree work (Spring 2002). Zhiyan Shao has done some of the preliminary work on the annotations and tags needed to implement this work within the IBM Websphere product.

References

[Borges 1999] Borges, J. A., Jimenez, J. and Rodríguez, N. J., "Speech Browsing the World Wide Web", IEEE International Conference on System, Man and Cybernetics, October 1999.

[Danielsen 2001] Danielsen, P. J. (2001) "The Promise of a Voice-Enabled Web", IEEE Computer, August 2000, Volume 33, Number 8, pp. 104-106.

[Goose 1998] Goose, S., Wynblatt, M., and Mollenhauer, H. (1998) "1-800-Hypertext: Browsing Hypertext With a Telephone" Proceedings of the Hypertext'98, Pittsburgh, PA, pp. 287-288.

[Lucas 2000] Lucas, B. (2000) "VoiceXML for Web-based distributed conversational applications" Communications of the ACM September 2000, Volume 43, Issue 9.

[Mabner 2000] Manber, U., Patel, A. and Robison, J. "Experience with Personalization on Yahoo!", Communications of the ACM, August 2000, V43, N8, pp. 35-39.

[Poon 2001] Poon, J., Nunn, C. (2001) "Browsing the Web from a Speech-based Interface" Proceedings of the Human- Computer Interaction - Interact '01, pp. 302-309.

[Perry 2001] Perry, M, O'hara, K., Sellen, A., Brown, B., and Harper, R. "Dealing with mobility ACM Transactions on Computer- Human Interaction (TOCHI) December 2001, Volume 8 Issue 4.

[Wang 2001] Wang, Q.Y., Shen, M.W., De Shi, R., and Su, H. (2001) "Detectability and Comprehensibility Study on Audio Hyperlinking Methods" Proceedings of the Human-Computer Interaction - Interact '01, pp. 310-317.

[Yankelovich 1996] Yankelovich, N. (1996). How do Users Know What To Say? ACM Interactions, Volume 3, Number 6, November/December 1996