WEBGALAXY – INTEGRATING SPOKEN LANGUAGE AND HYPERTEXT NAVIGATION[1]

Raymond Lau, Giovanni Flammia, Christine Pao, and Victor Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
http://www.sls.lcs.mit.edu, mailto:{raylau, flammia, pao, zue}@sls.lcs.mit.edu

ABSTRACT

The growth in the quantity of information and services offered online has been phenomenal. Nevertheless, access mechanisms have remained relatively primitive, requiring users to primarily point and click their way through a forest of Web links and to expend valuable cognitive capacities to track the geography of the Web space. Conversational systems can provide an intuitive, flexible multi-modal interface to online resources. The explosive growth of the World Wide Web, the continuing standardization of Web related technologies, and the growing penetration of Internet access enable us to embed a very thin client inside a standard Web browser, making conversational interfaces available to a much wider audience. This paper presents WebGALAXY, a conversational spoken language system for access to selected online resources from within a typical browser. A thin Java based client is employed as the front end with much of the speech and natural language processing occuring on remote servers.

1. INTRODUCTION

Fueled by the World Wide Web and Internet boom, there has been an enormous growth in the amount of resources available online. On the Web alone, there are presently over 31 million publicly accessible pages living in over 627 thousand servers and the numbers are growing literally every minute. While the availability of information continues to grow in mammoth propotions, the means of information access have remained relatively primitive. Point, click and type has remained the predominant user interface paradigm. The appearance of search engines has helped, but these tools are only capable of retrieving and displaying the information as is. The user is still relegated to exploring a maze of links and is forced to expend valuable and scarce cognitive capacities to track the geography of the net space.

We, as do others, believe that a speech interface to the Web is ideal, especially for naive users, because it is a natural, flexible, efficient and economical form of human communications. Some work of which we are aware in this area include:

Texas Instruments' SAM (speech aware multimedia, [3]) is a speech interface to a Web browser. The system extracts the text content from the hyperlinks in an HTML page and automatically and dynamically creates context free grammars and pronunciation graphs used by the speech recognizer running on the client machine, allowing users to speak aloud HTML page titles, bookmarks, hyperlinks, and mnemonic phrases for hotlists. Thus, speech becomes an alternative mode to the keyboard and the mouse in terms of Web navigation. The OGI SLAM system ([4]) is similar.

BBN SPIN ([11]) captures the user's speech via a plug-in for Windows95 and converts the audio into a series of VQ codewords. The coefficients are sent over the internet to a large vocabulary speech recognition server. The recognized sentence is returned and if the query is for one of a small number of domains like weather, the appropriate Web page is retrieved. Otherwise, the query is submitted to a search engine, AltaVista, whose results are displayed as if the query were typed.

However, we feel that providing a speech interface can go beyond providing for the ability to "speak" the links originally designed for keyboard and mouse. While such a capability is undoubtedly useful in hands-busy environments and for disabled users, it does not necessarily expand the system's capabilities nor lead to new user interaction paradigms. Instead, we should consider how we can expand a user's ability to obtain desired information easily and quickly. We view speech interfaces as augmenting, not replacing, the traditional mouse and keyboard. A user should be able to choose among many input/output modalities to achieve the task in the most natural and efficient manner.

When a user's request contains complex constraints or when the information space is broad and diverse, spoken language provides a particularly appropriate interface. Both of these situations occur frequently on the Web. For example, getting the annual report for a public company typically requires knowing its specific URL, clicking through multiple layers of links starting at the company's homepage (or perhaps a central repository's homepage) or using one of the keyword search engines hunting for the right keywords to use. The main problem here is that the traditional interfaces present only a limited set of choices at any point. Spoken language allows a user to sidestep the prescribed organization of the Web. Constraint specifications also occur frequently with services offered on the Web. These, though natural to users (e.g., seeking "a flight from Boston to Hong Kong with a stopover in Tokyo"), are are not easily covered by menu or form-based paradigms due to their rich structure. Spoken language, on the other hand, offers a user significantly

[1]This research was supported by DARPA under contract N66001-94-C-6040, monitored through Naval Command, Control and Ocean Surveillance Center.

```
{c identify
   :topic {q weather
             :quantifier def
             :pred {p in
                      :topic {q city
                                :name "boston" }  } }
   :subject 1
   :domain "Weather"
   :olang "english"
   :para "give me the weather report for Boston." }
```

Figure 1: Semantic frame for the request "What is the weather forecast for Boston?"

more power in expressing constraints, freeing them from rigid, preconceived indexing and command hierarchies.

The Web not only provides a natural application for conversational spoken language but it also provides a host of standards and technologies which makes it possible to bring spoken language systems to a much wider audience of users. For example, Java makes it possible to have a single platform independent client front end. The standardization of communications protocols (and hopefully soon, distributed object protocols) makes it possible for developers of spoken language systems to take advantage of a wide variety of Web based information and services without being encumbered by content production. Finally, Web protocols can also provide a lingua franca for communications between various components of a spoken language systems as well as with information sources. As an example, KTH has demonstrated their text-to-speech (TTS) system, which accepts requests via a Web forms protocol, with our WebGALAXY architecture.

2. GALAXY

The WebGALAXY system extends our GALAXY system ([2], [12]) to take advantage of the recent developments involving the World Wide Web and the Internet. GALAXY allows speech and natural language access to a variety of online information and services. A distributed client-server architecture is employed. A speech recognition server, using MIT's SUMMIT recognizer ([1]) with a 2300 word vocabulary, converts spoken input into an n-best list of hypotheses. These are then passed onto a natural language (NL) server, which uses MIT's TINA understanding system ([9]). User requests can also be entered by typing or by clicking, in which case the input is passed directly to the NL server.

The NL server returns a semantic frame, such as the one shown in Figure 1, with slots and values representing the user request. The frame is then forwarded to an appropriate information server to obtain the necessary information or to enter the appropriate transaction. Information servers communicate with a combination of local databases, HTTP, Gopher and SQL servers, proprietary commercial networks such as CompuServe, and potentially other online resources, to fulfill the user's request.

The information server returns a semantic frame that contains a response which may include a combination of: HTML output, a URL, or a natural language response. The GALAXY client is responsible for conveying the response to the user using the most appropriate modality (graphic display, synthesized speech), completing the current round of user interaction.

Information about the current state of the dialog is maintained between rounds to allow references to previous information (e.g., "Give me more information for the first one"). GALAXY currently handles requests about the following application domains:

Weather information about weather in numerous cities worldwide

Air Travel information about flights from American Airlines' easySABRE

City Guide information about various points of interest around the Boston area

Other domains under development within the GALAXY architecture include restaurant information ([10]) and automobile classifieds ([7]).

3. WebGALAXY

The original GALAXY system employed a client program running under the X Windows system on a workstation class computer. With the arrival of Web and Internet based technologies and standards, we realized that we can bring the spoken language technology of GALAXY to a much wider audience of users if we can relocate the user interface client to a standard Web browser. Today, anyone with an Internet connection and a Web browser can in theory use WebGALAXY and access information via any natural combination of speaking, typing, pointing and clicking. No additional software or plug-ins are required.

To implement WebGALAXY, we made several changes to the original GALAXY architecture. The previous GALAXY client's functionality was split into two parts: a new WebGALAXY hub and a standard Web browser. The hub maintains the state of the current discourse with the user and mediates the information flow between the various servers and the Web browser. The Web browser is used to provide all graphical user interface to WebGALAXY. Figure 2 outlines the WebGALAXY architecture.

We support two graphical user interfaces: a Java/JavaScript interface with rich interactivity and a forms interface. WebGALAXY is designed to also support a displayless interface, using only spoken language interaction. To start WebGALAXY, the user simply goes to the WebGALAXY homepage, selects an interface and indicates whether he would like to interact with the system via voice over a standard phone line or via typing and clicking only. Finally, she clicks the start button and a graphical user interface client is launched on her machine. If a phone number were provided, the user would be called as well. We describe the graphical interfaces next.

4. GRAPHICAL INTERFACES AND IMPLEMENTATION

The Java interface is the preferred graphical interface for WebGALAXY. An example display from the Java version is shown in Figure 3. The top area is where the Java applet resides. There is a status display ("Ready"), a box

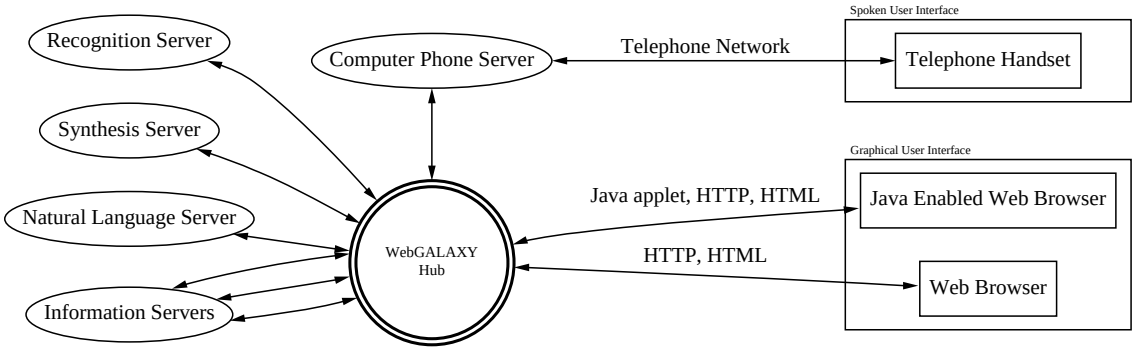

Figure 2: WebGALAXY Architecture

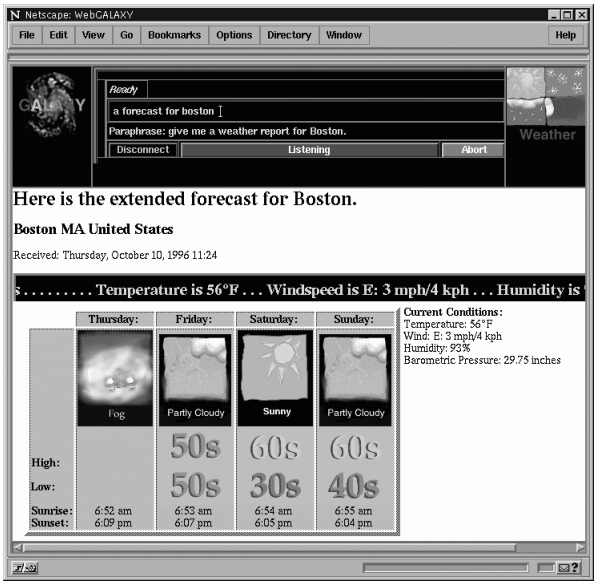

Figure 3: Java-based graphical interface

for either the recognized spoken input or the typed input (”a forecast for Boston”), a paraphrase for the parsed input (”give me a weather report for Boston.”), buttons for disconnecting from the system, aborting the current request, a combination button/status display for controlling and indicating the system’s listening state (”Listening”), and finally, an iconic indication of the domain of the last request (”Weather”). The lower portion of the browser window is used to display WebGALAXY’s response. For spoken input, either automatic endpoint detection or tap to talk modes can be used to detect when a person starts and finishes speaking. Audio tones and visual cues in the displayed listening status are provided to indicate when the system is listening to the user.

In the display shown here, the user asked for “the forecast for Boston” orally. It was misrecognized as “a forecast for boston” but the request was correctly handled by the natural language server. The reply with the forecast was generated by the Weather domain server and displayed by WebGALAXY. The user could have also typed the same request. Certain requests can generate lists. For example, the request “Show me Chinese restaurants in Cambridge” would generate a list as a reply. The user can then continue to interact verbally, using the names of the restaurants or their ordinal positions (“the second one”) or she can click on an item in the list and then say, for example, “Give me the phone number,” referring to the clicked item. The ability to support mixed-mode interactions allows the user to choose whatever input method is most convenient. For certain types of lists, clicking twice on an item gives more detailed information. Requests for homepages, such as “Show me the homepage for MIT” will retrieve and display the target homepage in the lower area. The user is free to continue browsing with the mouse and keyboard from that page, such as by clicking a link.

The Java interface utilitizes the full graphical display power of HTML supported by the browser. The applet itself performs a very limited but essential function by allowing the hub to push a new response output to the browser and also relays certain user requests, such as clicking an item in a list, back to the hub. We choose not to have it render the actual display output. While such an approach allows for precise control over the display, we feel that HTML is adequate for most displays and is capable of supporting particular customizations. However, for specialized applications, a custom Java controlled display may be appropriate (e.g., air travel in [5]). More details on our Java applet can be found in [6].

We have also implemented a forms version of WebGALAXY for use at locations where the resources do not support the Java version, e.g., where communications bandwidth is too low or where the browser does not support Java. The forms version is implemented by means of an added gateway that translates between the hub-applet API and standard HTML tags for forms. No changes to the hub or the remainder of the architecture are needed. More details about the forms implementation is available in [6]. We would like to note also that a displayless version of WebGALAXY can be implemented within the architectural framework just as easily.

5. CONCLUSION

WebGALAXY demonstrates that human language access to resources on the World Wide Web is feasible, albeit in limited application domains. The Internet is both a natural application for spoken language technology and is an en-

abler for widespread access to the technology. By taking advantage of developments in Internet related technologies, we can allow a much larger user base to benefit from conversational interfaces. The continued improvements in Internet telecommunications infrastructure allows us to permit users access to WebGALAXY via a very thin platform independent Java client, or a forms interface.

We have successfully tested WebGALAXY with Netscape Navigator 3.0 running under Windows, MacOS, Linux, SunOS, and Solaris, and with Microsoft Internet Explorer 3.0 running under Windows and from locations within the U.S., Europe and Asia, and with Internet connections as slow as 28,800 bps. We have also tested our systems behind security firewalls configured to still allow casual browsing of Web pages. Finally, we have experienced some initial success at interfacing our system with standard Internet telephony products. However, because of the experimental nature of the system, we are not ready to make WebGALAXY available to the general public.

We are witnessing a shift in the types of interfaces to the Web. Desktop browsers are being replaced by browsers that reside in many types of devices such as hand-held personal digital assistants, smart digital telephones, and television set-up boxes. Each one of these devices has specific input and output interfaces and limitations. A multi-modal user interface that supports typed and spoken natural language could provide easy and universal access to the Web from different devices and in multiple languages, reaching a much wider audience.

There is also a transformation in the type of semantic units on the web, from one of static modality (text and graphics) to multiple dynamic modalities (text and graphics, responses generated specifically for a request, speech, audio, and video data). All these diverse types of content and modalities require a paradigm shift in the content organization and underlying communication protocols. For example we are beginning to see work in organizing content for audio access ([8]).

WebGALAXY is a small step in the direction of shifting the paradigm of the user interface to the Web from a simple point-and-click navigation in a deep forest of HTML documents towards a richer, more flexible and intuitive navigation. WebGALAXY takes a naturally expressed request, fetches the information needed to respond, possibly from multiple online sources, and generates a summary response specific to the request. This is a fundamentally different paradigm than having the user access potentially several URLs from his bookmarks, traverse layers of links, and mentally compose a summary response. However, much work needs to be done on multiple fronts. From an engineering standpoint, we need to create authoring tools and architectural frameworks to support rapid application domain development. On the content side, we have already mentioned the need for content organizations to facilitate multi-modal access. We at the Spoken Language Systems group are also actively exploring multilingual access as well as displayess (speech only) access. Finally, WebGALAXY, and in general, spoken language access to the Web, would clearly benefit from advances in computer telephony integration, permitting the simultaneous transmission of data and voice input/output over the same connection line.

6. ACKNOWLEDGMENTS

GALAXY, upon which this work is based, represents the research efforts of many other current and former members of the Spoken Language Systems group, including E. Brill, J. Glass, D. Goddeau, L. Hetherington, E. Hurley, H. Meng, M. Phillips, J. Polifroni, and S. Seneff. We would also like to thank R. Schloming, S. Kwong, and S. Lee for their assistance with WebGALAXY.

7. REFERENCES

[1] J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-based speech recognition," in *Proc. ICSLP '96*, Philadelphia, PA, vol. 4, pp. 2277–2280, Oct. 1996.

[2] D. Goddeau, E. Brill, J. Glass, C. Pao, and M. Phillips, "Galaxy: A human-language interface to on-line travel information," in *Proc. ICSLP '94*, Yokohama, Japan, pp. 707–710, Sept. 1994. URL http://www.sls.lcs.mit.edu/ps/SLSps/icslp94/galaxy.ps.

[3] C. Hemphill and P. Thrift, "Surfing the web by voice," in *Proc. ACM Multimedia '95*, San Francisco, CA, pp. 215–221, Nov. 1995.

[4] D. House, "Spoken language access to multimedia (SLAM): A multimodal interface to the world-wide web," Master's thesis, Department of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, Beaverton, OR, Apr. 1995. Also appears as Technical Report TR 95-008 and URL ftp://speech.cse.ogi.edu/pub/docs/SLAM-thesis.ps.Z.

[5] L. Julia, L. Neumeyer, M. Charafeddine, A. Cheyer, and J. Dowding, "Http://www.speech.sri.com/demos/atis.html *(sic)*." Presented at AAAI '97 Spring Symposium, Mar. 1997.

[6] R. Lau, G. Flammia, C. Pao, and V. Zue, "WebGALAXY: Beyond point and click - a conversational interface to a browser," in *Proc. Sixth International World Wide Web Conference* (M. R. Genesereth and A. Patterson, eds.), Santa Clara, CA, pp. 119–128, Apr. 1997. URL http://www.sls.lcs.mit.edu/raylau/webgalaxy.

[7] H. Meng, S. Busayapongchai, J. Glass, D. Goddeau, L. Hetherington, E. Hurley, C. Pao, J. Polifroni, S. Seneff, and V. Zue, "WHEELS: A conversational system in the automobile classifieds domain," in *Proc. ICSLP '96*, Philadelphia, PA, vol. 1, pp. 542–545, Oct. 1996.

[8] T. Raman, "Cascaded speech style sheets," in *Proc. Sixth International World Wide Web Conference* (M. R. Genesereth and A. Patterson, eds.), Santa Clara, CA, pp. 109–117, Apr. 1997.

[9] S. Seneff, "TINA: A natural language system for spoken language applications," *Computational Linguistics*, vol. 18, no. 1, pp. 61–86, Mar. 1992.

[10] S. Seneff and J. Polifroni, "A new restaurant guide conversational system: Issues in rapid prototyping for specialized domains," in *Proc. ICSLP '96*, Philadelphia, PA, vol. 2, pp. 665–668, Oct. 1996.

[11] D. Stallard, "Demonstration of BBN SPIN (SPeech over the INternet)." Presented at MIT, Cambridge, MA (no published cite), 1997.

[12] V. Zue, "Navigating the information superhighway using spoken language interfaces," *IEEE Expert*, pp. 39–43, Oct. 1995.