A multimodal browser for the World-Wide Web

David G. Novick, David House, Mark Fanty, Ronald A. Cole
Center for Spoken Language Understanding
Oregon Graduate Institute
{novick, dhouse, fanty, cole}@cse.ogi.edu

Abstract

Spoken Language Access to Multimedia (SLAM) is a spoken language extension to the graphical user interface of the World-Wide Web browser Mosaic. SLAM uses the complementary modalities of spoken language and direct manipulation to improve the interface to the vast variety of information available on the Internet. To make the advantages of spoken language systems available to a wider audience, the speech recognition aspects can be performed remotely across a network. This paper describes the issues and architecture of what is believed to be the first spoken-language interface to the World-Wide Web to be easily implemented across platforms.

I. Introduction

The World-Wide Web (WWW) (CERN, 1994) is a network-based standard for hypermedia documents that combines documents prepared in Hypertext Markup Language (HTML) (NCSA, 1994a) with an extensible set of multimedia resources. The most popular browser for the WWW is Mosaic (NCSA, 1994b), a cross-platform program developed and distributed by NCSA, now running in X-based Unix, Macintosh and PC-Windows environments. As a hypermedia viewer, Mosaic combines the flexibility and navigability of hypermedia with multimedia outputs such as audio and GIF images. The World-Wide Web, especially viewed with Mosaic, is phenomenally popular. It is an archetypal interface to what will become the national information infrastructure. By mid-Spring of 1994, Internet traffic was doubling about every six months. Of this growth, the World-Wide Web's proportional usage was doubling about every four months. In absolute volume of traffic, use of the WWW was doubling every two and a half months (Wallach, 1994).

Much of the popularity of Mosaic can be attributed to its mouse-based interface, which can quickly, simply and directly aid the user in browsing a vast variety of documents on the Internet. However, inherent limitations in mouse-based interfaces make it difficult for users to perform complex commands and to access documents that cannot be reached by the visible links. Speech-based interfaces, on the other hand, perform well on these types of complex, nonvisual tasks.

In this paper, we discuss the complementary nature of mouse- and speech-based interfaces. We then present the Spoken Language Access to Multimedia (SLAM) system, which adds speech recognition to Mosaic. We describe a method of distributed processing of speech understanding across a network by passing speech to a receiving results from a remote recognizer. Finally, we examine the current status of the SLAM project and outline directions for future research.

II. Interface modalities for hypermedia

The graphical user interface, especially with pointer-based direct manipulation, has become the predominant model for human-computer interaction. Even in innovative settings such as the World-Wide Web, which provides a rich hypermedia environment that includes outputs in hypertext, images and sound, the inputs to the system remain keyboard- and pointer-based. (As the most typical pointer is the mouse, we will use the term "mouse-based" interface to refer to pointer-based interfaces generally.)

The mouse-based direct-manipulation interface (Shneiderman, 1983) provided a rational and innovative means of interaction with computer systems. While physical pointing (and bitmapped displays) solved many of the problems with character-and-keyboard-based interfaces, direct manipulation based on physical pointing did not make use of the full range of expressive capabilities of human users. This omission was, no doubt, mostly a consequence of the relatively poor state of other means of expression as input modalities; spoken-language systems have made immense progress in the ten years since 1983 (Cole, Hirschman et al., 1995).

If the technology permits, adding spoken-language capabilities to hypermedia holds the promise of extending users' abilities in ways they find appealing. Empirical studies of multimodal interfaces have looked at user preferences for different kinds of inputs. For example, Rudnicky (1993) showed that users preferred speech input, even if it meant spending a longer time on the task, when compared with using a keyboard and a direct manipulation device. Oviatt and Olsen (1994) found that users of multimodal interfaces displayed patterns of use that reflected the contrastive functionality of the available modalities.

Other researchers have investigated the comparative advantages of multimodal interfaces, including Cohen (1992) and Oviatt (1992, 1994). One of the goals of this research has been to attempt "to use the strengths of one modality to overcome for the weaknesses of another" (Cohen, 1992, p. 143), who proposed a framework for this analysis. Natural language systems overcome some of the weaknesses of pointer-based interfaces by allowing the specification of context, temporal relations, and unseen objects. On the other hand, language has the problem that the user may not know the vocabulary of the recognizer. Spoken language systems are also prone to other problems such as ambiguity and other recognition errors (Cohen, 1992).

A. Mouse-based interfaces to hypermedia: advantages and disadvantages

The physical pointing involved in mouse-based interfaces is the source of both advantages and disadvantages for this modality. From the user's perspective, pointing has the traditional advantage of direct manipulation, namely reference specified deictically and implicitly through a combination of action and reference, as in double-clicking an icon to start a program. Moreover, normal practices in programming GUI interfaces result in the provision of immediate feedback to the user that the reference was successful, typically by highlighting the selected entity. From the point of view of the author of a WWW document, mouse-based pointing has the advantage that the reference can be completely specified: the label of a link will appear exactly as the author wrote it. Additionally, physical pointing in this context has no referential ambiguity; when the user clicks a mouse button, the user and the author both know exactly to which entity the user is referring.

Mouse-based interfaces also have a number of disadvantages, particularly of the "lost-in-hyperspace" variety. This well-known problem was identified for hypertext systems by Whalen and Patrick (1989), who proposed a text-based natural-language solution. We suggest that is because when reference is based on physical pointing to a graphically-represented entity, the absence of such an entity on the screen means that the user *cannot* refer to it. In other words, the act of reference depends on the physical location of the referent's presentation, which in hypermedia may be pages and documents away.

Hypermedia interfaces typically have standard features such as a "hot list" and history windows in order to give users a place that contains references they might want and that are otherwise not displayed. But the user might also prefer to refer to an entity by a name other than that specified by the author; the only way the user has to specify an entity is to click on it. Finally, the spatial nature of the interface limits the set of things to which the user can refer. Users cannot *describe* entities (Cohen, 1992) instead of pointing. Similar problems exist with respect to actions. Because actions are typically accomplished by selecting a command from a menu or by clicking on an icon, it is difficult to express complex actions other than as a perhaps tedious series of primitives. The advantages and disadvantages of mouse-based interfaces for hypermedia are summarized in Table 1.

Advantages	**Disadvantages**
1. Deictic reference and combination of action and reference	1. Reference depends on location of referent
2. Author completely specifies the representation of the entity	2. (a) User might prefer another representation and (b) no other representation possible
3. No referential ambiguity	3. Vocabulary of references limited to those with visible links
4. Generally gives immediate feedback that user's action was understood	4. Difficult to express complex acts

Table 1: Mouse-Based Interaction with Hypermedia

B. Spoken-language-based interfaces to hypermedia: advantages and disadvantages

Fortunately, many of the advantages and disadvantages of spoken-language-based interfaces for hypermedia turn out to be complements of those for mouse-based interfaces. From the user's standpoint, the ability to refer to an entity no longer depends on the location of its graphical representation. Indeed, *all* referents are potentially available because the user can say the name of the referent without having to see it displayed. A related advantage is that the user can now have a number of different ways in which to refer to entities. Similarly, multiple action primitives could easily be combined into a single complex action which could include temporal and other sophisticated concepts that are not expressible in mouse-based interfaces. Other advantages of spoken-language input to hypermedia include the freeing of the user's hands for other activities. Indeed, it might be possible to build a spoken-language-only interface to hypermedia that could serve users by telephone instead of requiring a GUI.

Speech input to hypermedia also has characteristic disadvantages, which are often reciprocal consequences of its advantages. For example, because references no longer depend on physical location, references may become ambiguous: a "hot link" may be uniquely accessible via the mouse but ambiguously accessible via speech because another hot link might have the same label. This factor strongly suggests that designers of hypermedia interfaces should avoid multiple uses of "Click here" as a hot-link label; rather they should use a lexically meaningful label that refers to the semantics of the linked entity. Similarly, there could be confusion between names of labels and names of actions.

Although speech interfaces make all referents available, the user may not know all available referents. However, the user is no worse off than in the mouse-based case, where it is not even possible to refer to other entities directly.

Although the hands-free nature of spoken-language interfaces is appealing, early implementations of spoken-language interfaces to hypermedia may have to rely on "push-to-talk" methods so that the recognizer is not confused by extraneous speech (Lunati & Rudnicky, 1990). Similarly, while spoken-language understanding could possibly provide a speech-only interface, there would be a number of problems with unimodal application of speech to hypermedia, including (a) straining user tolerance in getting through extended synthesis of text, (b) loss of meaning from images, (c) difficulty in navigation, and (d) not immediately knowing the names of new links. Indeed, consideration of these factors suggests that the application of spoken-language technology as a multimodal extension to a hypermedia browser would likely be more immediately useful than development of a unimodal, speech-only interface. Even in a multimodal interface, there remain open issues. How, for example, could a user use spoken language to refer to a bitmap or other image? The advantages and disadvantages of spoken-language input for hypermedia are summarized in Table 2.

The major conclusion we draw from this comparison is that mouse-based and speech-based modalities have a high degree of complementarity that could improve the usefulness of hypermedia systems. This could lead to a *synergistic* interaction style (Lefebvre et al., 1993; Nigay & Coutaz, 1993) that allows multiple modalities to perform a task.

Advantages	**Disadvantages**
1. References no longer depend on location	1. Possible ambiguity
2. All referents are available	2. User may not know all available referents
3. Hands free	3. Might have to use touch-to-talk to avoid extraneous sounds and speech

Table 2: Spoken-Language Interaction with Hypermedia

Advantages	Disadvantages
4. Could provide access to information when GUI not available	4. Problems with audio-only: (a) too much text (b) pictures (c) navigation (d) presentation of links
5. More direct expression	5. Unknown words, unlimited vocabularies
6. More than one way to refer to an entity	6. Multiple links may have same key words; or link and command may be the same
7. Can express more complex action	7. Difficult to refer to graphics such as bitmaps, icons and pictures

Table 2: Spoken-Language Interaction with Hypermedia

III. Project description

SLAM adds spoken language as an input to the Mosaic browser by enabling interaction with a remote server that provides (a) speech-capable documents and (b) the recognition systems needed to use them. SLAM provides a relatively simple extension to Mosaic plus access to a SLAM server at OGI that performs speech recognition for a set of speech-capable hypermedia documents. In these documents, users will are able to select hot-links with spoken language. A major advantage of this approach is that recognizers will not have to be developed and supported for all client platforms; rather, spoken-language interaction can be added to additional platforms through creation of new versions of the SLAM-extended Mosaic, which will not require major modification of Mosaic.

SLAM is the first generally available spoken-language interface to the World-Wide Web that could be easily implemented across platforms. No other such interface has been reported in the literature. If a spoken-language interface is used in the “workaday world” of cooperative computing (Moran, 1990) exemplified by the Web, then we will have (a) empirical evidence of its utility and (b) a fund of varied experiences with the interface that could contribute to improvements. From a practical standpoint, the idea is to make the interface available and see what happens—as in the case of the original Mosaic interface and other WWW browsers.

A. SLAM architecture

SLAM is a spoken-language interface system for Mosaic based on local access to a remote recognition-capable Web server. The overall architecture, as depicted in Figure 1, is based on a Web server that has spoken-language software and “speech-capable” HTML Web documents. The SLAM server receives, recognizes and responds to requests from users running an extended ver-

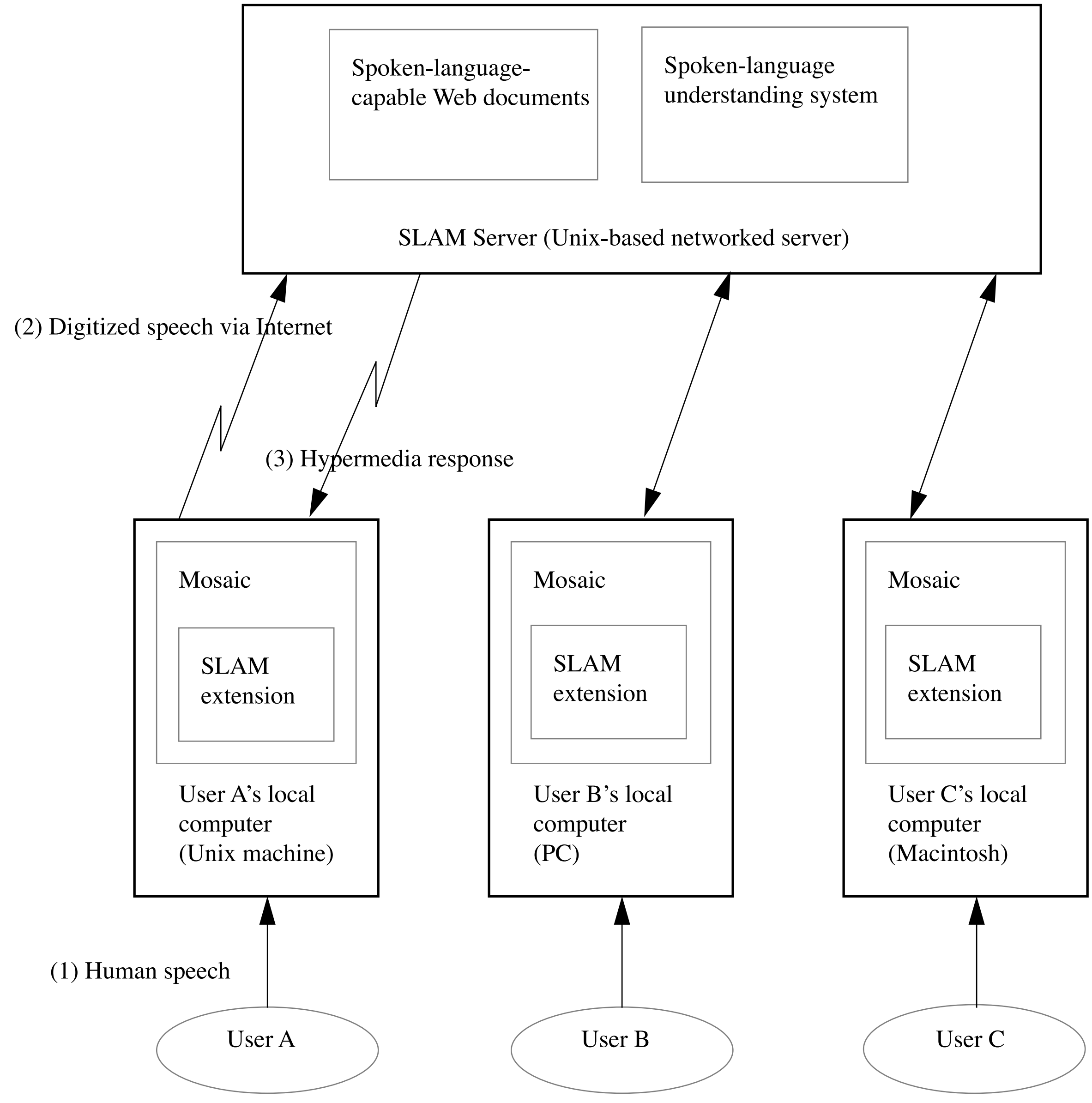
Spoken-language-capable Web documents
Spoken-language understanding system
SLAM Server (Unix-based networked server)
(2) Digitized speech via Internet
(3) Hypermedia response
Mosaic
SLAM extension
User A's local computer (Unix machine)
Mosaic
SLAM extension
User B's local computer (PC)
Mosaic
SLAM extension
User C's local computer (Macintosh)
(1) Human speech
User A
User B
User C

Figure 1. SLAM Architecture

sion of Mosaic on their local computer. Users can use spoken language to select hot-links in documents on the SLAM server.

Users on heterogeneous platforms—such as Macintoshes, PC's and X Windows interfaces for Unix—will interact as usual with their local Mosaic browser to the World-Wide Web. As indicated by Arrow 1 in Figure 1, the user speaks to his or her local machine. The local Mosaic browser will contain extended code that digitizes the user's utterance and, as indicated by Arrow 2, sends the digitized signal to the SLAM server. The server processes the speech signal and matches the utterance to a WWW uniform resource locator (URL). As indicated by Arrow 3, the server then sends back to the local machine a hypermedia response, typically a new HTML document.

As this discussion indicates, SLAM's architecture is based on a client-server model where the local browser does not (necessarily) perform recognition and the remote server provides both speech-capable documents and the speech recognition necessary for their full use. We call this approach the remote-recognition model. The alternative, called the local-recognition model, would require the local browser client to provide speech-recognition capabilities.

The remote-recognition model provides a number of advantages. First, it will help to spread the popularity and use of spoken-language systems without the hardware costs otherwise associated with such systems. Because the recognition is being done remotely, the user could use a relatively inexpensive machine with limited memory and still perform WWW navigation. Second, this approach could also serve as a foundation for a speech-only interface over the telephone, which would allow the user to access the variety of useful information available over the Internet without needing a terminal. Other advantages of the remote-recognition model include being able to control and collect the spoken utterances of the users from around the world for the building of standard language corpora, which will lead to further research in the field. Also, as the state of the art in speech-recognition capabilities improves, the software would only need to be updated at the central SLAM server site instead of at all sites that were using the interface.

One possible disadvantage of the remote-recognition model is that the transfer rate to and from the central recognizer may be quite slow; however, given the likely short length of the transferred speech and the normal delays in accessing WWW documents anyway, this effect does not appear to be serious. Another possible problem with having a central server is the risk of additional delays arising from multiple clients trying to access the single recognition server simultaneously. A final consideration is that SLAM product will not allow access directly to the worldwide network of Internet documents; authors of HTML documents will have to prepare speech-capable documents specially, or eventually, have a script to automatically create speech-ready versions of existing documents.

B. SLAM implementation

The principal components of or implementation of SLAM include a minor extension to Mosaic's GUI and the networking and recognition modules associated with the server. The major functions operate as follows:

1. As the SLAM system starts up, the user's "hot list" information file consisting of links stored from previous sessions is read in from a file in the user's home directory. SLAM also saves pronunciation models for the "hot list" items in the user's home directory, so that these models do not need to be generated on the fly.
2. The user can use the extended browser to navigate the WWW in the same way that they use the Mosaic browser, by using the mouse to select hot-links within the current document to bring up other documents.
3. Once the user reaches a speech-capable document (denoted by an icon at the top of the document), the user is also able to use the touch-to-talk speech facility of SLAM to select

links. For documents that are too long to fit entirely on the current browser screen, SLAM views the document in its entirety, rather than focusing only on the part of the page which is visible to the user. This enables the user to use speech to specify items that do not appear on the current page.

4. When the speech button is pressed, three things happen:

   a. Mosaic sends to the server the URL of the current document, so that the server can set up the recognizer with the right vocabulary.

   b. Mosaic prepares to accept a new document from the server in the usual manner, except that as the document comes back its headers are parsed for the speech pragmas.

   c. A SLAM function digitizes speech from the systems' usual audio input and sends it to the open SLAM server.

5. The SLAM server compares the speech to the possible results which came from the user's "hot list" and current page links, and returns the URL corresponding to the result back to the client machine. The client's extended Mosaic browser then retrieves the document specified by the URL.

The current version of SLAM does not handle grammars of spoken inputs, but rather only recognizes "hot list" phrases and phrases relating to link labels. In the future, SLAM will handle grammars so that a much greater variety of inputs may be used. In doing so, many of the advantages which spoken language have over direct manipulation can become implemented in our system. For example, use of anaphoric references, reference to multiple documents and actions, and specification of temporal events will all be possible with the application of grammars to the system.

The speech recognizer determines the appropriate target vocabulary and phrases through the SPEECH= tags, the arguments of which is are pronunciations models of the label. In the current implementation of SLAM, SPEECH= tags appear together at the top of the "speech-ready" in a form such as:

```
<SPEECH= ao r eh g ax n [.pau] g r ae j uw ih t [.pau] ih n s t ih t uw t

                 w eh dh axr >
```

in which each line in the SPEECH= section relates to a corresponding link label within the file (Note: [.pau] refers to a "pause" in the speech). For example, the first pronunciation model within this example, relating to the words "Oregon Graduate Institute," would relate to the first link in that document, which would be of a form like:

```
<A HREF="http://www.ogi.edu/"> Oregon Graduate Institute</A>
```

while the second pronunciation model would relate the link corresponding to the word "weather."

C. SLAM speech recognition

OGI has used neural-network-based recognition for limited vocabulary tasks for a number of years (Cole et al., 1990; Fanty et al., 1993). SLAM uses the OGI general-purpose recognizer described by Cole et al. (1994). Vocabulary independence is necessary because of the high number of people's names and other proper nouns which compose labels for hypertext links. The ability to generate accurate phoneme representations of these labels in near real-time would be a valuable step towards developing a future system which does not rely on "speech-ready" documents. SLAM uses a context-independent, task-independent phonetic classifier trained on the OGI Continuous English Speech Corpus, which contains the unconstrained speech of 690 speakers, each talking for up to one minute.

SLAM uses a dictionary to find word pronunciations, and uses automatic text-to-phoneme mapping to create pronunciations for words not in the dictionary. With the current system, "speech-ready" pages are created semiautomatically by processing each link label first through the Moby dictionary (which is strictly a table lookup) and through the Orator text-to-speech system (which can create highly accurate pronunciation models even for words the system has never seen).

This method is slower but more accurate than other methods, which is acceptable because often "speech-ready" pages will be generated off-line and pronunciation model accuracy will be a major criterion for in the system's successful performance. However, in future systems in which the user can visit any WWW page with speech, text-to-phoneme transformations will need to be performed in near-real-time.

A wide variety of microphones and recording environments which will be used by remote users of SLAM. This sort of variation typically has a significant impact on the accuracy of recognizers. Hynek Hermansky of OGI and Nelson Morgan of ICSI have developed RASTA spectral processing for robustness to different recording environments (Hermansky et al., 1994). Thus SLAM will use RASTA instead of PLP to increase recognition robustness.

D. Further challenges

One problem not directly addressed by our system is that of non-English hypertext labels. As this is the *World-Wide* Web, links can appear in a variety of languages and not all of the sounds from these languages can be mapped to English phonemes. A short-term solution to this problem would be to map these sounds to their closest English equivalents or to use the flexibility of speech interfaces to map these link names to words which *do* have corresponding English equivalents, if possible. A longer-term solution would be use non-English language corpora (for which SLAM may be used to aid in collecting as international users send their speech to a central recognizer to perform recognition) to train the recognizer on non-English as well as English phonemes.

Currently SLAM does not allow for the selection of highlighted pictures and icons, although with Mosaic you can click on these items to bring up other WWW pages. By using the filename relating to the icon or picture, as well as labels within the HTML code to give a descriptive name to the image, it may be possible to accurately specify these images in future versions of SLAM. A further challenge in this area will be the selection of parts of items known as imagemaps, which in Mosaic call different WWW pages depending on *where* within the image one clicks. This would seem to be a very difficult task to accomplish in general with speech alone and may be one task which is better restricted to multimodal or other systems which used mouse-based input.

IV. Conclusion

SLAM is serving its designed function as an exploratory system of multimodal access to multimedia. In particular, the ability to access off-screen referents (such as items much further down on the current page or different "hot list" pages) with speech when used in conjunction with Mosaic's original ability to quickly refer to any on-screen item has show the complementary modalities can be used to compensate for each others' weaknesses.

We will soon complete work on the SLAM server which will allow for remote recognition of speech input to the WWW browser, which will allow systems which do not currently have speech recognition capabilities to still be able to take advantage of this spoken language enhancement to Mosaic. This pioneering effort is a step towards the eventual goals of an unconstrained multimodal interface to the WWW, as well as providing some foundations for a speech-only system for accessing information from the WWW.

For more information, visit the SLAM Web page at http://www.cse.ogi.edu/SLAM/.

Acknowledgments

The authors thank Oscar Garcia for his contributions to the remote-recognition model. This research was supported by National Science Foundation award number IS-9422461.

References

CERN (European Laboratory for Particle Physics) (1994). "World-Wide Web Home", URL: http://info.cern.ch/, undated.

Cohen, P. (1992). The role of natural language in a multimodal interface, *Proceedings of UIST'92*, 143-149.

Cole, R., M. Fanty, Y. Muthusamy and M. Gopalakrishnan (1990). Speaker-independent recognition of spoken English letters, *International Joint Conference on Neural Networks*, San Diego, CA, June 1990, v.2, 45-51.

Cole, R., Hirschman, L., et al. (1995). The challenge of spoken language systems: Research directions for the Nineties. *IEEE Transactions on Speech and Audio Processing*, 3(1), 1-20.

Cole, R., Novick, D., Burnett, D., Hansen, B., Sutton, S. & Fanty, M. (1994). Towards automatic collection of the U.S. Census, *Proceedings of ICASSP* '94, Adelaide, South Australia, April 1994, v.1, 93-96.

Fanty, M.,Schmid, P., & Cole, R. (1993). City name recognition over the telephone, *Proceedings of ICASSP* '94, Minneapolis, April 1993, v.1, 549-552.

Hermansky, H. (1990). ''Perceptual Linear Predictive (PLP) Analysis of Speech,'' *J. Acoust. Soc. Am.*, v. 87, no 4.

Hermanky, H., Morgan, N.& Hirsch, H. (1994). Recognition of Speech in Additive and Convolutional Noise Based on RASTA Spectral Processing, *Proceedings of ICASSP* '94, Adelaide, South Australia, April 1994, v.1, 421-424.

Lunati, J.-M., & Rudnicky, A. (1990). The design of a spoken language interface. *Proceedings of the DARPA Speech and Natural Language Workshop*, Hidden Valley, PA, June 1990, 225-229.

Lefebvre, P., Duncan, G., & Poirier, F. (1993). Speaking with computers: A multimodal approach. *Proceedings of EuroSpeech'93*, Berlin, 1665-1668.

Moran, T., & Anderson, R. (1990). The workaday world as a paradigm for CSCW design. *Proceedings of CSCW'90*, Los Angeles, CA, October 1990, 381-393.

Nigay, L., & Coutaz, J. (1993). A design space for multimodal systems: concurrent processing and data fusion. *Proceedings of InterCHI'93*, Amsterdam, April, 1993, 172-178.

NCSA (National Center for Supercomputing Applications) (1994a). "A Beginner's Guide to HTML," URL: http://www.ncsa.uiuc.edu/General/Internet/WWW/HTMLPrimer.html (undated).

NCSA (National Center for Supercomputing Applications) (1994b). "NCSA Mosaic Home Page," URL: http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/NCSAMosaicHome.html, undated.

Oviatt, S. (1992). Pen/voice: Complementary multimodal communication. In *Proceedings of Speech Tech'92*, New York, February 1992, 238-241.

Oviatt, S., & Olsen, E. (1994). Integration themes in multimodal human-computer interaction. *Proceedings of ICSLP'94*, Yokohama, 551-554.

Rudnicky, A. (1993). Factors affecting choice of speech over keyboard and mouse in a simple data-retrieval task. *Proceedings of EuroSpeech'93*, Berlin, 2161-2164.

Shneiderman, B. (1983). Direct manipulation:A step beyond programming languages. *IEEE Computer*, 16(8), 57-69.

Wallach, D. (1994). "WWW Size," World-Wide Web page summarizing NSFNet statistics collected at nic.merit.edu, URL: http://www.cs.princeton.edu/grad/dwallach/www-talk/size.html, March 22, 1994.

Whalen, T., & Patrick, A. (1989). Conversational hypertext: Information access through natural language dialogues with computers. In *Proceedings of CHI'89*, 289-292.