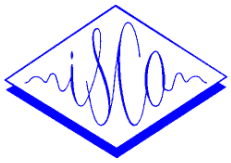

ESCA Workshop on Interactive
Dialogue in Multi-Modal Systems (IDS-99)
Kloster Irsee, Germany
June 22-25, 1999

# A System Design for Human Factors Studies of Speech-Enabled Web Browsing

L.J. Adams, R.I. Damper, S. Harnad and W. Hall
Department of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK

## Abstract

This paper describes the design of a system which will subsequently be used as the basis of a range of empirical studies aimed at discovering how best to harness speech recognition capabilities in multimodal multimedia computing. Initial work focuses on speech-enabled browsing of the World Wide Web, which was never designed for such use. System design is complete, and is being evaluated via usability testing.

## 1 Introduction

Speech technology has now advanced to the stage where it offers great promise for human-computer interaction in a variety of applications [1, 2]. Applications have to be chosen and engineered very carefully, however, with human factors given full consideration, if real gains are to be achieved [3, 4, 5]. In particular, the early, popular belief that speech was somehow a 'universal' medium – better in all respects than all other media – is too simplistic. Attention in the research community is turning toward the optimal deployment of speech I/O in multimodal interfaces [4, 5, 6, 7]. Key findings are that speech-only interfaces have some problems but that users have a strong preference for interacting multimodally [8]. At the same time, speech is notable for its absence in current multimedia systems [9, 10]. Yet Furui [11] – one of the most respected speech scientists in the world – has commented:

> "Input by speech recognition and output by text and graphics is an ideal combination in most interactive systems ..." and: "Dialogue modelling in a multimedia environment is a very new, important and interesting research topic."

Speech offers unique advantages over more conventional media. For instance, the concept of the every-citizen interface in the US [12] rests on the realisation that a large proportion of the population lacks the computer literacy (or even the linguistic capability) to use conventional, text-based I/O. Also, speech offers the user another I/O channel in a complex, multimodal interactive system, so effecting a classical separation of modalities in terms of multiple-resource theory [6, 13, 14, 15, 16]. Oviatt [7, 8] provides some detailed information on how this capability can be utilised in practical systems: since speech is a poor modality for defining spatial position, a pointing device is used to specify *where* (in terms of screen location) something is to be done, while a concurrent speech command specifies *what* is to be done. Thus, it is a foregone conclusion that hypermedia and multimedia system capabilities will be integrated with user capabilities in speech production and perception in the not too distant future.

This integration is currently held back by limited understanding of human-computer spoken interaction. In the words of Oviatt [7]: "Inadequate research from this perspective has left a gap in our scientific knowledge, hindering our ability to support robust speech for real commercial applications". The ultimate goal is an all-purpose system with which one can interact conversationally as we do with one another. But it is partly a cognitive question – calling for experimental research on people's interactions with physical systems – how speech capability can best be integrated with future multimedia computing of the kind envisaged by Harnad [17] and used in conjunction with other input/output (I/O) modalities. For many easily imaginable applications (e.g. interacting with a system in the dark, while one's hands are otherwise occupied, or if one is handicapped or at a remote site), direct speech control over what we otherwise control by keyboard or mouse is essential. But even where keyboard or mouse would be accessible, their limited fit to human capabilities means that it may be more natural and comfortable to interact conversationally, by speech. In other cases, such as operating on real or virtual 2-D or 3-D objects, or parallel spatial arrays of graphical or textual information, keyboard or mouse may *become* the interactive tools of choice once the relevant tactile, interactive skills have been mastered. Yet during the early stages of use – from complete unfamiliarity, through the first interactive steps, via trial-and-error, systematic instruction and example, until independence and mastery have been achieved – speech interaction may nevertheless be the op-

timal modality for everyone (cf. the "every-citizen" interface [12]). Hence, a key issue is how speech is best used within a multimodal, multimedia system.

## 2 Speech-Enabled Web Browsing

As a first step in answering the questions identified above, we have designed and implemented a speech-enabled system which will subsequently be used as the basis of a range of empirical human factors studies. Initial work focuses on browsing of the World Wide Web [18].

The advantages of speech almost certainly depend on the degree of *constraint* [19] imposed on users by the system and/or application. Loosely speaking, constraint is inversely related to size of response set. The conventional means of navigating the Web provides an example of a high-constraint interface. Here the user is presented with a document in which the author or the system has highlighted links to other documents or applications: these represent the main or sole possibilities for navigation. An example of a low-constraint application would be the more query-based interaction with a hypermedia system, as embodied by generic and compute-link facilities (e.g. MICROCOSM [20, 21]), and where the user requests more information about a topic that may or may not be highlighted as a button (e.g. "Tell me more about *elephants*"). We hypothesise that speech will become increasingly useful as the domain becomes less constrained. Thus, we intend eventually to study human performance factors with both high- and low-constraint interfaces and to apply the results to subsequent development, in the spirit of user-centered, 'participatory' design [22]. For the moment, however, we will concentrate on high-constraint Web browsing, based on Microsoft Internet Explorer version 4 (MSIE4) using object linking and embeddding (OLE).

## 3 System Design

Current Web browsers were not designed to accept spoken commands nor were they designed to facilitate human factors experimentation. Hence, significant technical effort must be expended in system design and implementation for our purposes. Many of the problems which arise are novel and challenging: they relate to assumptions made by implementors regarding the way their software products will be used and which do not anticipate our requirements.

The experimental system is based on the IBM ViaVoice speech recogniser (Version 4.1) – a large-vocabulary, continuous speech, speaker-dependent device. Software is written in Visual C++ (version 5.0), to exploit the well-known advantages of the object-oriented paradigm and for compatibility with the speech application programming interface (API), provided by ViaVoice (which is written in C). The application handles speech input by using specific calls to the SMAPI Speech Manager provided by ViaVoice. The software system is rather large. Hence, the Rational Rose CASE (computer aided software engineering) tool is used to manage program complexity automatically. Details of this CASE tool can be found at:

`http://www.rational.co.jp/products`

The complete software system is divided into three modules: Interface Design; Basic Navigation; and Enhanced Navigation. A formal (context-free) grammar in Backus-Naur form is used to define the command syntax. This allows for easy modification and automatic generation of the executable grammar module. The grammar is deliberately restricted to reduce the cognitive load placed on the user while learning to use the system.

The system design uses four diagrammatic object-oriented representations, using the Unified Modelling Language (UML) methodology [23]. The modelling representations used are use cases, class diagram, state diagram and interaction diagrams. The class diagram (figure 1) gives a conceptual overview of the system.

## 4 Navigation by Speech

The goal was to allow the user to control by speech a subset of the navigation facilities offered by MSIE4. Specifically, these were:

- basic toolbar commands – *back*, *forward*, *stop* and *refresh*.
- URL loading – using a form of 'spoken bookmark'.
- URL spelling – allowing the user to compose a spoken bookmark.
- following hyperlinks – see below.
- page printing.

Given the space restrictions of this paper, we limit ourselves here to describing perhaps the most important feature – the ability to follow hyperlinks by speech. The intention is that users simply speak the command *follow* or *tell me about* followed by the word(s) on-screen associated with the hyperlink. The browser then loads the destination location (URL) and retrieves the document. However, the software implementation was fraught with problems resulting from the wide variety of HTML document layouts, parsing the documents for hyperlink extraction, 'cleaning' the hyperlink text by removal of spurious and out-of-vocabulary text, lack of support in the MSIE4 object module for spoken interaction, etc.

Access directly to hyperlinks through the MSIE4 object model or to the raw HTML document for hyperlink

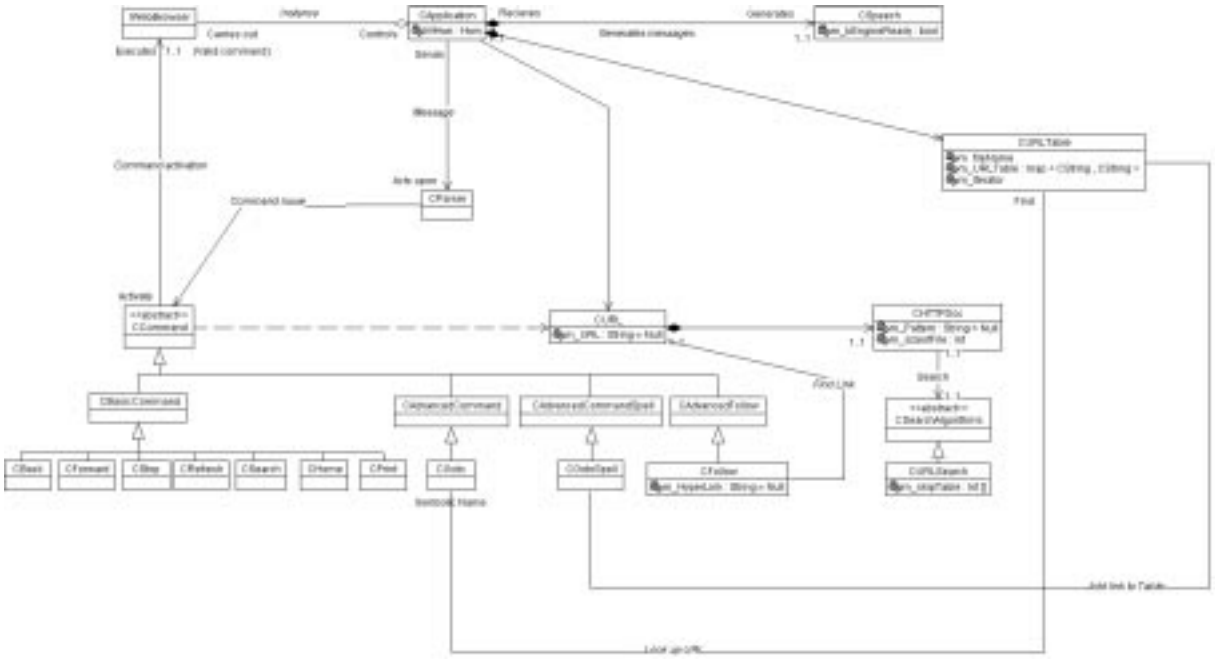

Figure 1: Class diagram for the speech-enabled Web browser, illustrating the conceptual design of the system.

searching were required but proved difficult. After much effort, the former approach was abandoned. The final implemented solution uses thread programming for background document retrieval and processing. During application initialisation, an event timer is instantiated to cause a function to be invoked once per second. This function then determines if the current URL has been changed. If so, a new thread of execution initiates page retrieval ready for subsequent processing. The page is then stored in an hash table, so that it remains available for browsing without reprocessing/reloading.

The speech-enabled Web browser has been extensively tested by white-box testing through the course of implementation and by black-box testing at the end [23].

## 5 Usability Testing

This has only just started – there is much more to be done. At this stage, four subjects only (including the first author) have participated in informal testing. After enrolment *on the first author's voice*, subjects were then asked to enter the basic commands (five), exercising the basic functionality of the browser. After a short period of familiarisation, these commands were recognised with greater than 95grammar quickly and easily.

Subsequently, the composition of spoken-bookmarks and following-hyperlinks functions were exercised. This naturally led to many more errors (recognition accuracy falling to something like 60from a tightly-controlled command grammar to arbitrary word recognition on arbitrary strings (which don't conform to English *n*-gram word grammar). Also, the ViaVoice is not truly speaker-independent. Once bookmarks are trained as single words or short phrases, however, recognition accuracy improves dramatically.

The main strategy for error recovery is user confirmation, the level of which is in user control by clicking in a dialog box. This proved to be a useful and popular approach. The URL spelling facility was found to be awkward and time-consuming. It was widely disliked with participants preferring to enter arbitrary text on the keyboard.

Generally, subjects felt that speech would take time to get used to, but eventually would make a valuable addition to keyboard/mouse input devices, allowing them to perform the more simple functions without distraction (e.g. swapping hand position from keyboard to mouse).

## 6 Conclusions

A speech-enabled Web browsing system has been described, based on IBM's ViaVoice recogniser and Microsoft's Internet Explorer. Many technical problems had to solved and many remain: these arise because the World Wide Web has not so far evolved in a way which anticipates and supports speech interaction. A small, informal usability study has been undertaken. This reveals that speech has potential for future incorporation into a multimodal interface in which users can exploit its unique properties when appropriate and ignore it when not. Future work will concentrate on more formal evaluations to confirm these initial findings.

# References


[1] P. R. Cohen and S. L. Oviatt. The role of voice input for human-machine communication. *Proceedings of the National Academy of Sciences, USA*, 92:9921–9927, 1995.

[2] C. Kamm, M. Walker, and L. Rabiner. The role of speech processing in human-computer intelligent communication. *Speech Communication*, 23:263–278, 1997.

[3] A. F. Newell. Speech – the natural modality for man-machine interaction? In B. Shackel, editor, *Human-Computer Interaction – INTERACT '84*, pages 231–235. Elsevier (North-Holland), Amsterdam, 1985.

[4] R. I. Damper. Speech as an interface medium: How can it best be used? In C. Baber and J. Noyes, editors, *Interactive Speech Technology: Human Factors Issues in the Application of Speech Input/Output to Computers*, pages 59–71. Taylor and Francis, London, 1993.

[5] R. A. Sharman. Speech interfaces for computer systems: Problems and potential. *Displays*, 14:21–31, 1993.

[6] G. L. Martin. The utility of speech input in user-computer interfaces. *International Journal of Man-Machine Studies*, 30:355–375, 1989.

[7] S. Oviatt. User-centered modeling for spoken language and multimodal interfaces. *IEEE Multimedia*, 3(4):26–35, 1996.

[8] S. Oviatt. Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction*, 12(1–2):93–129, 1997.

[9] M. G. Helander. Foreword. In C. Baber and J. Noyes, editors, *Interactive Speech Technology: Human Factors Issues in the Application of Speech Input/Output to Computers*, pages ix–xii. Taylor and Francis, London, 1993.

[10] R. I. Damper. Foreword. In R. I. Damper, W. Hall, and J. W. Richards, editors, *Multimedia Technologies and Future Applications*. Pentech Press, London, 1994.

[11] S. Furui. Prospects for spoken dialogue systems in a multimedia environment. In *Proceedings of European Speech Communication Association (ESCA) Tutorial and Research Workshop on Spoken Dialogue Systems: Theories and Applications*, pages 9–16, Vigsø, Denmark, 1995.

[12] B. Tognazzini. Ordinary citizens and the national information infrastructure. In *Proceedings of National Research Council Workshop: Towards an Every-Citizen Interface to the National Information Infrastructure*, Washington, DC, 1996.

[13] R. A. North. Task functional demands as factors in dual task performance. In *Proceedings of the Human Factors Society 21st Annual Meeting*, pages 367–371, San Antonio, TX, 1977.

[14] C. D. Wickens, D. L. Sandry, and M. Vidulich. Compatibility and resource competition between modalities of input, central processing and output. *Human Factors*, 25:227–248, 1983.

[15] R. I. Damper, A. D. Lambourne, and D. P. Guy. Speech input as an adjunct to keyboard entry in television subtitling. In B. Shackel, editor, *Human-Computer Interaction – INTERACT '84*, pages 203–208. Elsevier (North-Holland), Amsterdam, 1985.

[16] R. I. Damper, M. A. Tranchant, and S. M. Lewis. Speech versus keying in command and control: Effect of concurrent tasking. *International Journal of Human-Computer Studies*, 45:337–348, 1996.

[17] S. Harnad. Interactive cognition: Exploring the potential of electronic quote/commenting. In B. Gorayska and J. L. Mey, editors, *Cognitive Technology: In Search of the Humane Interface*, pages 397–414. Elsevier, Amsterdam, 1995.

[18] 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 & Technology, Portland, OR, 1995.

[19] K. S. Hone and C. Baber. Modelling the effects of constraint upon speech-based human-computer interaction. *International Journal of Human-Computer Studies*, 50(1):85–107, 1999.

[20] W. Hall. The role of hypermedia in multimedia information systems. *ACM Computing Surveys*, 27(4):599–601, 1995.

[21] L. Carr, D. de Roure, H. Davis, and W. Hall. Implementing an open link service for the World Wide Web. *World Wide Web Journal*, 1(2):61–71, 1998.

[22] S. Kuhn and M. J. Muller. Participatory design. *Communications of the ACM*, 36:25–103, 1993.

[23] R. S. Pressman. *Software Engineering: A Practioner's Approach (4th Edition)*. McGraw-Hill, New York, NY, 1997.