

# **Let's stop pushing the envelope and start addressing it: a reference task agenda for HCI**

**Steve Whittaker, Loren Terveen and Bonnie A. Nardi**  
**ATT Labs-Research**

## **Running head: Reference task agenda**

### **Abstract**

We identify a problem with the *process of research* in the HCI community – an overemphasis on “radical invention” at the price of achieving a common research focus. Without such a focus, it is difficult to build on previous work, to compare different interaction techniques objectively, and to make progress in developing theory. These problems at the research level have implications for practice, too; as researchers we often are unable to give principled design advice to builders of new systems. We propose that the HCI community try to achieve a common focus around the notion of *reference tasks*. We offer arguments for the advantages of this approach, as well as considering potential difficulties. We explain how reference tasks have been highly effective in focussing research into information retrieval and speech recognition. We discuss what factors have to be considered in selecting HCI reference tasks and present an example reference task (for searching speech archives). We conclude with recommendations about necessary steps to execute the reference task research agenda, including both required technical research, as well as changes in HCI research community practice. The technical research involves: identification of important user tasks by systematic requirements gathering; definition and operationalisation of reference tasks and evaluation metrics; execution of task-based evaluation along with judicious use of field trials. Perhaps more important, we also suggest changes in HCI community practice. We must create forums for discussion of common tasks and methods by which people can compare systems and techniques. Only through this can the notion of reference tasks be integrated into the process of research and development, enabling the field to achieve the focus it desperately needs.

## 1. THE PROBLEMS WITH HCI AS RADICAL INVENTION

Research in HCI, particularly as embodied in the CHI conference, focusses largely on novel problems and solutions that push the technology envelope. Most publications describe novel techniques or novel applications of existing techniques. Newman (1994) provides quantitative evidence for this. He compared CHI with five other engineering research fields, such as thermodynamics and aerodynamics, using content analysis to classify abstracts of published papers to identify their research contribution. In other engineering disciplines, over 90% of published research built on prior work, contributing: (1) better modelling techniques (allowing predictions about designs); (2) better solutions (addressing previously insoluble problems); and (3) better tools and methods (to apply models or build prototypes). However, only about 30% of CHI papers fit into these cumulative categories. The majority either reported “radical” solutions (new paradigms, techniques, or applications) or described experience and heuristics relating to radical solutions.

### 1.1 Radical Invention is not always Effective

This analysis strongly suggests that CHI differs from other engineering research disciplines. We offer arguments that the current state of affairs is problematic with respect to two different success criteria.

One criterion consistent with radical invention is *technology transfer*. One motivation for constant innovation is the example of whole new industries being created by novel user interfaces. Applications like Visicalc and Lotus 123 drove the early PC market, and Mosaic/Netscape led to the Web explosion. In this view, HCI research is an engine room from which novel interaction techniques are snatched by waiting technology companies. There undoubtedly are some success stories according to this criterion, including collaborative filtering (Communications of the ACM 1997, Goldberg et al., 1992, Hill et al., 1995, Resnick et al., 1994, Shardanand & Maes 1995), UI toolkits and general programming techniques (Rudishill et al., 1996). The early graphical user interfaces developed at Xerox PARC (Smith et al., 1982) successfully combined ideas such as overlapping windows and the mouse that predated the coalescence of the HCI community. These ideas then made their way into the Macintosh and Microsoft Windows systems.

Nevertheless, user interfaces with widespread impact generally originated *outside the HCI community* (Isaacs & Tang, 1996). Visicalc was invented by a business student and a programmer. CAD systems developed from Sketchpad (Sutherland, 1963), and were independently invented by engineers at Boeing and General Motors (Foundyler, 1984). AOL and Instant Messenger were invented by business people. Tim Berners-Lee, the inventor of HTML and the Web, while a computer scientist, was not a member of the CHI community.

The second success criterion is scientific. The radical invention model has not aided the development of a *science* of HCI. This is a controversial area with acrimonious past debate concerning the scientific basis of HCI (Newell & Card,

1985, Carroll & Campbell, 1986), and extended arguments about the relationship of HCI to psychology and cognitive science. There are isolated pockets of HCI research deriving basic precepts from psychological theories (Card et al., 1983, Gray et al., 1993, Olson & Olson, 1990). However, these papers are in the minority (as is evident from Newman's analysis), and they may not have major effects on mainstream HCI practice (Landauer, 1995, Newman, 1994). The analysis so far should make it clear why this is so. Consolidation is impossible if everyone constantly is striking off in novel directions. While radical invention is vital to making progress, so too is cumulative research. Concepts must be clarified, tradeoffs determined, key user tasks and requirements described, metrics or critical parameters (Newman, 1997) identified, and modelling techniques constructed. We are simply not doing enough of this type of work.

## **1.2 What we don't Know: Requirements, Metrics and Uses of Everyday Technologies**

One significant problem originating from the absence of cumulative research is the lack of clear understanding of *core* user tasks, interactive technologies and techniques. We lack systematic information about tasks that are essential to people's everyday computing activities: browsing, retrieval and management of Web information; use of email and voicemail; personal information management; and task management<sup>1</sup>. While there are many radical solution attempts in these areas, we lack accepted bodies of knowledge about these everyday computer activities. In many of these areas, while a few initial studies have been conducted, there is no consensus about user tasks, no common view of outstanding issues and problems, and no accepted success metrics. Thus, when addressing these problems, researchers must begin by carrying out their own research to identify requirements, and evaluation metrics. This difficulty is manifest for information retrieval (Amento et al., 1999, Whittaker et al., 1998a), asynchronous communication (Whittaker & Sidner, 1996, Whittaker, Davis, Hirschberg & Muller, 2000), and desktop UIs (Barreau & Nardi, 1995). The absence of shared task information makes it difficult to focus research problems, to compare research results, and determine when a new solution is *better*, rather than simply *different* (Newman, 1997).

A well-known problem with radical invention is that it often is not based on an understanding of user tasks and requirements. Researchers thus find themselves proposing radical solutions to problems that are of little interest to users, while neglecting genuine problems. Barreau and Nardi (1995) studied how users organised desktop information. Most people felt that their computer files were adequately organised, and that archiving tasks did not require major support. Nevertheless, much recent technological work has addressed archival support

---

<sup>1</sup> By systematic bodies of knowledge, we employ the *very weak criterion* that at least two studies have been conducted in a given area. Note that we are not even insisting that the studies agree on their core findings. There are often one or two pioneering studies in a given domain, after which no further research is done.

(Fertig et al., 1996, Gifford et al., 1991, Rao et al., 1994). On the other hand, many people experienced problems in transferring information between applications. Here basic empirical investigation uncovered an important task that was not being addressed by the research community. This insight led to work on Apple Data Detectors (Nardi, Miller & Wright, 1998), now a part of the Macintosh operating system. The research also identified a second requirement that desktop organisers should support, namely *reminding*. Users remembered outstanding tasks by simply inspecting folders and files. This research thus discovered two novel user problems, (and hence criteria for evaluating new versions of desktop organisers), as well as finding that a commonly addressed problem, archiving, actually didn't deserve as much attention.

In addition to a lack of shared task and requirements descriptions, we also have little systematic data about how people use popular technologies. We lack information about how people actually use email, voicemail, cellular phones, the Windows interface, digital personal organisers, and instant messaging<sup>2</sup>. The popularity of these technologies and their widespread usage make it imperative to know how people use them, what they use them for, how successful they are, and where problems lie.

Furthermore, we don't have a good understanding of *why* certain core user interface techniques are successful. GUIs are central to the enterprise of HCI, and although we have successful guidelines for building them (Shneiderman, 1982), we lack theoretical understanding of why they are successful (Baecker, 1987, Brennan, 1990).

And of course, new radical innovations such as immersive virtual realities, augmented realities, affective computing, and tangible computing make the problem worse. Not only do we not understand these new technologies and their basic operation, we don't have a clear sense of how much innovation is tolerable or desirable. In sum, although we lack basic understandings of current users, tasks and technologies, the field is encouraged to try out even more radical solutions, without pausing to do the analysis and investigation required to gain systematic understanding.

### **1.3 How we don't Know it: the Dissemination Problem**

Furthermore, even when a useful body of knowledge exists for a core task, the HCI community does not have institutions and procedures for exploiting this knowledge. We advocate workshops for articulating and disseminating knowledge of core tasks and practices. Changes in community standards, e.g., reviewing guidelines for the CHI conference, and in HCI instruction, are also needed for new practices to take hold. These will allow our suggestions to be institutionalised.

---

<sup>2</sup> One complicating factor is that some proprietary research has been conducted into these technologies in industrial contexts. Nevertheless we still need publicly available data about technologies used by millions of people multiple times a day.

## 2. THE REFERENCE TASK SOLUTION

To address the overemphasis on radical invention and lack of knowledge about important tasks, we propose a modified methodology for HCI research and practice centred on *reference tasks*. Our proposal has both technical and social practice aspects. We discuss: (1) how reference tasks may be represented and used, and (2) new practices that the HCI community must adopt in order to develop and utilise reference tasks.

The goal of reference tasks is to capture and share knowledge and focus attention on common problems. By working on common tasks central to HCI, the community will enjoy these benefits:

- Shared problem definitions, datasets, experimental tasks, user requirements and contextual information about usage situations will allow greater research focus;
- Agreement about metrics (e.g., Newman's (1997) critical parameters) for measuring how well an artifact serves its purpose enables researchers and designers to compare different user interface techniques objectively, and to determine when progress is being made;
- Advice to designers will be based on a stronger foundation, namely knowledge about core tasks within a domain and the best techniques for supporting the tasks;
- Theory development also will be strengthened; the relationship between core tasks, interface techniques and critical parameters provides the basis for a predictive model.

Our proposal partly overlaps Roberts and Moran (1983) and Newman (1997). Roberts and Moran (1983) argue for the use of standard tasks in evaluating word processing applications. Our proposal differs in being independent of a specific application. Newman suggested using critical parameters to focus design on factors that made critical differences to user interface performance. We are motivated by Newman's original findings (1994) and wish to underscore the importance of critical parameters. However, we offer a broader approach that emphasises the relationship between requirements, reference tasks and metrics. Newman's account is unclear about the methods by which critical parameters are chosen. Another concern is that metrics may be task-specific rather than general - as his approach would seem to imply. Finally, we address the institutional processes required for the approach to work, in particular, how researchers can jointly identify reference tasks, collect data, analyze tasks, disseminate and make use of shared results.

### 2.1 Reference Tasks in other Disciplines

To motivate our approach, we trace the role of related concepts in speech recognition and information retrieval.

#### **Speech Recognition (The DARPA Workshops)**

Until the late 1980s, speech recognition research suffered from the same problems as HCI research. Researchers focussed on different tasks and datasets, making it

difficult to compare techniques and measure progress. Then, DARPA organised an annual workshop where researchers meet for a “bakeoff” to compare system performance on a shared dataset. (Marcus, 1992, Price, 1991, Stern, 1990, Wayne, 1989). A dataset consists of a publicly available corpus of spoken sentences, divided into training and test sentences. The initial task was to recognise the individual sentences in the corpus. There was no dialogue, and there were no real-time constraints. The success metric was the number of correctly recognised words in the corpus.

At each workshop, participating groups present and analyse their system performance. The utility of different techniques can thus be quantified – identifying which techniques succeed with certain types of data, utterances or recognition tasks. All interested researchers get an annual snapshot of what is working, what isn’t, and the overall amount of progress the field is making.

And progress has indeed been made. Initial systems recognised small vocabularies (1000 words), had response times of minutes to hours, and high error rates (10%). Current systems recognise much larger vocabularies (100,000 words), operate in real-time, and maintain the same error rate while recognising increasingly complex spoken sentences. Furthermore, as system performance improves, more difficult tasks have been added to the bakeoff set. Early corpora consisted of high quality audio monologues, whereas more recent tasks include telephone quality dialogues. More recent developments include attempts to extend these methods to interactive tasks (Walker et al., 1998).

Shared datasets have other benefits independent of the annual bakeoffs. There are now standard ways to report results of research taking place outside bakeoffs. Independent studies now report word error rates and performance in terms of shared datasets, allowing direct comparison with known systems and techniques.

### **Information Retrieval (The TREC Conferences)**

A core set of tasks and shared data have also successfully driven research in Information Retrieval. The Text REtrieval Conference (TREC), (Voorhees & Harman, 1997, 1998) sponsored by the United States National Institute of Standards and Technology (NIST), is analogous to DARPA speech recognition workshops.

A major goal of TREC is to facilitate cross-system comparisons. The conference began in 1991, again organised as a bakeoff, with about 40 systems tackling two common tasks. These were *routing* (standing queries put to a changing database, similar to news clipping services), and *ad hoc queries* (similar to search engine queries). Metrics for evaluation included *precision* – the proportion of documents a system retrieves that are relevant and *recall* – the proportion of all relevant documents that are retrieved. More refined metrics, such as average precision (for multiple queries at a standard level of recall), also are used.

The field has made major progress over 7 years: average precision has doubled from 20% to 40%. Furthermore the set of TREC tasks is being refined and expanded beyond routing and ad hoc queries. Over the years new tasks have been added, such as interactive retrieval, filtering, Chinese, Spanish, cross-lingual, high precision, very large collections, speech, and database merging. In each case,

participants address a common task with a shared dataset. Common tasks and metrics have made it possible not only to compare the techniques used by different systems, but also to compare the evolution of the same system over time (Sparck-Jones, 1998).

Similar approaches have been applied successfully in other disciplines such as digital libraries and machine learning.

### 3. REFERENCE TASKS IN HCI

#### 3.1 Lessons from DARPA and TREC

The case studies using shared tasks, metrics, and datasets reveal a number of relevant lessons. First, there are a number of positive outcomes:

- They show the essential role of the research community. Researchers defined tasks, produced and shared datasets, and agreed on suitable evaluation metrics. Furthermore, community practices were changed. Groups applied their systems to common tasks and data, then met to present and analyze their results. The bakeoff became a key community event.
- The basic task set is continuously refined. Both sets of workshops have added more tasks, increasing task difficulty and realism. This suggests that discovering “ideal” reference tasks will be an iterative collective process.
- One unexpected outcome is that system architectures and algorithms have become more similar. In consequence, it has become possible to carry out independent “black-box” evaluations of different modules. In the case of IR, this common architecture has also become a de facto decomposition of the overall retrieval task.
- A common architecture and shared datasets allows wider participation. Small research groups can evaluate their techniques on a sub-part of the overall task, without needing to construct complete experimental systems.

Several more problematic issues also arise:

- The workshops rely on a *bakeoff model*, assuming that research results are embodied in working systems that can be evaluated according to objective metrics. But how well will the system-bakeoff model work for HCI?
- Are there key HCI results that cannot be implemented, and thus cannot be evaluated as part of a system? Are there alternatives to the bakeoff model? Might we extend the bakeoff model to areas of HCI that are not focussed on systems, e.g., design, methods or requirements analysis? For methods, does ethnomethodological analysis yields better design data than an experiment? When are different methods useful (Gray & Saltzman, 1998)? Furthermore, the bakeoff itself is not strictly necessary although it serves an important social function. We can distinguish different elements of the DARPA/NIST process; shared datasets could be provided without bakeoffs to compare performance. Obviously this would decrease social interaction surrounding the meetings, but it would still allow for direct system comparison.
- There are also complex issues concerning *interactivity*. TREC and DARPA

have focussed mainly on non-interactive tasks. Going from simple tasks (with definable objective metrics) to more difficult and realistic tasks is not straightforward. Doing it may require fundamentally different algorithms and techniques. Both workshops have found difficulty in moving towards interactive tasks with subjective evaluation criteria.

- Previous evaluations allowed researchers to test systems on existing datasets, enabling the calculation of objective success measures such as word error rate, precision and recall. Bringing humans into the evaluation (as users, subjects, judges) produces a more complicated, costly, and subjective process. If HCI wants to experiment with the bakeoff model, it must begin precisely where other workshops have experienced problems.
- We previously interpreted system convergence positively, but it also may have a negative side. In both workshops, groups sometimes take the strategy of imitating the best system from the previous bakeoff, with additional engineering to improve performance. If this strategy is generally followed, the overall effect is to reduce research diversity, which may mean that techniques do not generalise well to novel problems. It is therefore critical that reference tasks sets are continually modified and made more complex to prevent “overlearning” of specific datasets and tasks.

We do not yet have solutions for these issues. Instead, we view them as cautions that must be kept in mind as we experiment with reference tasks.

### **Criteria for Selecting Reference Tasks**

How then do we choose appropriate reference tasks for HCI? Candidate reference tasks need to be *important* in everyday practice. A task may be “important” for different reasons:

- *Real* – first, tasks must also be “*real*”, that is, not divorced from actual user practice.
- *Frequent* - a task should be central to multiple user activities, so that addressing it will have general benefits. An example here might be processing asynchronous messages. Given the centrality of communication for many user activities, improved ways to manage messages will have widespread benefits;
- *Critical* - other tasks may occur less frequently, yet require near-perfect execution. Safety critical applications such as air traffic control are the prime example.

These criteria cannot be determined by researchers’ intuitions: significant empirical investigations of user activity are needed. We believe the following areas are worthy of intense study and are likely to yield reference tasks:

- information browsing, retrieval, and management;
- task management;
- information sharing;
- computer mediated communication;
- document processing;
- image processing and management;
- financial computation.

In selecting reference tasks, we also must avoid obsolescence. While radical inventions cannot be anticipated, we should exclude tasks that may become unimportant, or be transformed radically through predictable technological progress.

Our goals in defining reference tasks include generating shared requirements, accepted task definitions, descriptive vocabulary, task decomposition, and metrics. Common task definitions are critical for researchers to determine how other research is related to their effort. We will discuss how reference tasks are to be defined, and give an illustrative example. First, however, we think it is worthwhile to discuss potential drawbacks of our approach.

### **Potential Objections to our Proposal**

One potential problem is that HCI research may shift from innovation to become merely a “clean up” operation, directed solely at improving existing tasks, techniques, and applications. However, the areas of information retrieval and speech recognition provide hopeful counter-examples. Developments in speech recognition have led to successful applications to novel and important problems such as searching speech and video archives – and TREC has added tasks in these areas (Voorhees & Harman, 1997, 1998).

Furthermore, a shift away from innovation may be necessary: the history of science and technology indicates that many major inventions required a critical mass of innovators producing multiple versions of a given technology before its successful uptake (Marvin, 1988). Working in a radical invention mode precisely fails to achieve critical mass and thus the repeated solution attempts needed for adoption. Again, we are not declaring a moratorium on radical invention, just arguing for a different emphasis – HCI needs more “normal science” is needed (Kuhn, 1996).

There is also the danger of adopting a faulty paradigm. Progress in a field is severely limited when it is based on commonly accepted assumptions, but these assumptions are flawed. Cognitive Science and Artificial Intelligence have seen much lively debate over foundational assumptions (Dreyfus 1992, Ford & Pylyshyn 1995, Harnad 1990, Searle 1981). The notion of representation that was taken for granted in symbolic AI has been attacked (Bickhard & Terveen 1995). Similar arguments have been offered in the speech community. When non-interactive tasks and the sole performance metric of word error rate were central, techniques based on Hidden Markov models were popular. However, these techniques do not generalise well to “non-standard” situations such as hyperarticulation (Oviatt, 1996) or speech in noisy environments (Junqua, 1999). We do not believe the reference task approach runs this risk, however. Instead of proposing new assumptions or a new theory, we are suggesting a modified methodology, with more attention being paid to existing tasks. And note that completely radical *solutions* are consistent with our approach; they just need to be made relevant to a reference task and followed up by systematic analysis. We need a more rigorous understanding of the core conceptual territory of HCI so that we can better understand the role of radical innovations.

A variant of this last argument is that reference tasks induce bias towards the quantifiable, and a concurrent blindness to more subtle considerations. Much recent HCI work has shown how factors that are not easily quantified, such as ethical issues (Nardi et al., 1996) and social relationships among various stakeholders (Grudin, 1988, Orlikowski, 1992), affect the success of interactive technologies. From a design perspective, aesthetic issues also have a substantial impact on the success of applications (Laurel, 1990). Nevertheless, the reference task approach is neutral with respect to such factors. Insofar as factors are crucial to user performance and satisfaction in a given task, successful reference task definitions naturally must incorporate them. Many of these issues also may relate to *subjective user judgements*. Our later discussion on metrics addresses the role of subjective measures such as user satisfaction. Our hope is to discover systematic ways that users make decisions about interfaces. By defining appropriate methods to elicit this information, we can address this problem.

#### **4. HOW TO DEFINE A REFERENCE TASK**

We adopt the activity theory view that a task is a conscious action subordinate to an object (Kaptelinin, 1996). Each action, or task, supports some specific object such as completing a research paper, building an aeroplane or curing a patient. The object in these cases is the paper, the sale, the aeroplane, the patient. The tasks are performed to transform the object to a desired state (complete paper, closed sale, functioning aeroplane, healthy patient).

The same tasks can occur across different objects; thus, the task of outlining, may be useful in writing a book, preparing legal boilerplate, or specifying a product. In studying reference tasks it is important to determine the object of tasks so that appropriate customisations can be offered. While there might be a generic “outlining engine,” outlining a product specification could entail special actions that require customising the basic engine. Keeping the object in mind will bring designs closer to users requirements.

We also need empirical work to determine good domains for investigating candidate reference tasks. Of the many tasks involving computers, we must identify tasks satisfying our criteria of frequency and criticality. Defining a reference task may begin with an analysis of prior relevant work. All too often, each individual research effort defines its own problem, requirements, and (post-hoc) evaluation metrics. However, by analysing a broad set of related papers, one can abstract common elements:

- What are the user requirements in this area? Are they based on solid empirical investigation? Often the answer is no – which means more empirical studies of user activity are needed.
- Is there a common user task (or set of tasks) that is being addressed?
- What are the components of the task(s)? Is a task decomposition given, or can one be abstracted from various papers?
- What is the range of potential solution techniques? What problems do they address, and what problems are unsolved? Are there problems in applying

various techniques (do they require significant user input, scaling, privacy or security concerns)?

- How are solution techniques evaluated? Are metrics proposed that generalise beyond the single originating study? This last issue is crucial – it captures Newman’s (1997) “critical parameters” that define the artifact’s purpose and measure how well it serves that purpose.

If researchers abstract tasks from related work, they may be personally satisfied with the result. But other researchers may have different perspectives on all task aspects. For this reason, important community practices need to be introduced. Representative researchers and practitioners concerned with a particular area need to meet to discuss, modify, and approve the reference task definition. This would be like a standards committee meeting, although faster and more lightweight. Such groups might meet in the CHI workshops programme or in government sponsored workshops, organized by NIST or DARPA, for example. After a reference task is approved, its definition would be published, e.g. in the *SIGCHI Bulletin* and *Interactions*, with the complete definition appearing on the Web. But agreed reference task definitions also need to be modifiable, as researchers and practitioners experiment with them. One might use the NIST TREC model in where tasks are discussed annually, with modifications being made in the light of feedback.

Finally, the community must reinforce the important role of the shared knowledge embodied in reference tasks. Educational courses must show how tasks are defined, and the benefits from using this knowledge, as well as emphasising the problems that the reference task approach addresses. And the CHI review process could be modified so that reviewers explicitly rate papers with reference to the reference task model.

## **5. AN EXAMPLE REFERENCE TASK: BROWSING AND RETRIEVAL IN SPEECH ARCHIVES**

We now discuss an example reference task: browsing and retrieval in speech archives. It is intended to illustrate (a) identifying reference tasks; (b) using them to evaluate and improve user interfaces, and (c) issues arising in this endeavour. We summarise work reported in recent research papers (Choi et al., 1998, Nakatani, et al., 1998, Whittaker et al., 1998a, Whittaker et al., 1998b, Whittaker et al., 1998c, Whittaker et al., 1999, Whittaker et al., 2000). Other areas would have served equally well in illustrating reference tasks; we selected this area simply because of our personal expertise in this domain.

### **5.1 Selecting and Specifying Reference Tasks in the Domain of Speech Archives**

Two criteria we proposed earlier for selecting a reference task were that the task is either *frequent* or *critical*. So what is the evidence that accessing speech data is an important user task? Conversational speech is both frequent and central to many everyday workplace tasks (Chapanis, 1975, Kraut et al., 1993, Whittaker et al.,

1994). Voice messaging is a pervasive technology at work and at home, with both voicemail and answering machines requiring access to stored speech data. In the US alone, there are over 63 million voicemail users. New areas of speech archiving are also emerging, with television and radio programs becoming available on-line. These observations indicate that searching and browsing speech data meet the criteria of being frequent, general and real. Furthermore, we will show that the tasks we identify in speech retrieval generalise to retrieval of textual data, making it possible to use them more widely.

But identifying the *area* of speech retrieval does not identify specific user tasks when accessing speech archives. We therefore collected several different types of data concerning people's processing of voicemail. We chose to examine voicemail access rather than other audio data such as news, because voicemail is currently the most pervasive speech access application. We collected qualitative and quantitative data for a typical voicemail system, Audix<sup>TM</sup>: (a) server logs; (b) surveys from high volume users; (c) interviews with high volume users. We also carried out laboratory tests to confirm our findings on further users.

We found three core speech access tasks: (a) *search*, (b) *information extraction*, and (c) message *summarisation*. *Search* is involved in *prioritising* incoming new messages, and for *locating* valuable saved messages. Our working definition of search is: given a set of messages, identify a (small) subset of messages having relevant attributes with certain values (for example being from a particular person or being about a particular topic). *Information extraction* involves accessing information from *within* messages. This is often a laborious process involving repeatedly listening to a message for verbatim facts such as caller's name and phone number. Our definition of information extraction is: given a (set of) message(s) and a set of relevant attributes, identify the values associated with those attributes. A final task at the message level is *summarisation*: to avoid repeatedly replaying messages users attempt to summarise their contents, usually by taking handwritten notes, consisting of a sentence or two describing the main point of the message. We define summarisation as involving selection of a subset of information from within the document that best captures the meaning of the entire document. For more formal definitions of summarisation we refer the reader to Sparck-Jones (1998).

These three tasks were generated by analysis of voicemail user data. Nevertheless, although they originated from *speech* data, we found analogues in the independently generated TREC *textual* retrieval tasks. The fact that these three tasks are common to searching both speech and text is encouraging for the reference task approach. It argues that there may be general search tasks that are independent of media type.

*Figure 1 about here*

## 5.2 Defining Metrics

Our data also suggested possible metrics for gauging task success. In the interviews, people oriented to three different aspects of system usage when trying to execute their tasks. First, users wanted to complete their tasks correctly and

accurately: people repeatedly accessed voicemail until they had correctly extracted critical information, or until they had located a relevant message. We call this criterion *task success*. But people were also focused on *efficiency*: most users complained that executing the three core tasks was tedious, requiring too many user actions. This led to the metric of *task completion time* (Burkhart, Hemphill & Jones, 1994, Newman, 1997). Finally users made comments about the experiential quality of the interaction, leading to the criterion of *subjective evaluation*.

### 5.3 Task Oriented Evaluation of a Speech Browsing System

We next applied these task definitions and metrics to a real system that allows users to search and browse recorded news broadcasts<sup>3</sup>. The system applies automatic speech recognition to recorded broadcasts, indexes the resulting errorful<sup>4</sup> textual transcriptions for information retrieval and provides a user interface to support search and browsing (for a full architectural description, see Choi et al., 1998). Figure 1 shows the UI, which is described elsewhere (Whittaker et al., 1998c, Whittaker et al., 1999). The elements of the UI support a new paradigm for speech retrieval interfaces: “*What you see is (almost) what you hear*” (WYSIAWYH).

To evaluate two different versions of the UI (and hence two different UI techniques) we conducted laboratory experiments where users were given three tasks: *search*, *summarisation*, and *information extraction*, corresponding to the three reference tasks we had identified. For search, users had to find the most relevant speech document addressing a given issue. For summarisation, they had to produce a 6-8 sentence summary of a single speech “document” (where documents were about 5 minutes in length). For information extraction, they had to find a fact in a given speech document (*What were the names of the actors who starred in the Broadway musical “Maggie Flynn”?*). We used three evaluation metrics: *task success*, *time to solution*, and *perceived utility* of the user interface. To determine task success we had independent judges rank documents for relevance, rate summaries, and determine the correctness of factual answers.

*Figure 2 about here.*

We initially used the method to compare two different versions of the user interface. The main problem with browsing speech is that of random access to relevant materials. When browsing *text* people are able to visually scan exploiting structural cues (formatting, paragraphs, headers) to look for key words, enabling focus on relevant document regions. One version of the UI attempted to emulate this by providing a visual analogue to the underlying speech allowing people to visually scan as they would with text (see Figure 1). This WYSIAWYH UI provided users with graphical information about how the terms in their query were distributed in a given document, allowing them to “zoom in” on regions containing large numbers of query terms, and ignore the parts of the document

<sup>3</sup> We are also currently carrying out similar experiments on voicemail data (Whittaker et al., 2000).

<sup>4</sup> The errors arise because ASR performance for this type of data is imperfect: the state of the art is that about 70% of words are correctly recognised.

that weren't relevant to their query. It also provided information about the content of each speech document by presenting the errorful transcript of each story (including highlighted query terms) allowing users to visually scan through stories to identify relevant regions for playing. We compared this with a version of the UI without these browsing features. It allowed users to search for speech documents, but provided no browsing support: users selected audio to play using tape-recorder type controls (see Fig 2). For all metrics, the more complex UI was better for *search* and *information extraction* tasks, but we observed no differences between UI versions for summarisation. More details are supplied in (Whittaker et al., 1999).

We have since conducted further studies using identical metrics but subsets of the task set to evaluate different versions of the UI, and also the effects of systematically varying the quality of automatic speech recognition on browsing and search. We found that improving ASR quality beyond 84% accuracy made no difference to performance although users could detect subjective differences qualitatively (Stark et al., 2000).

#### **5.4 General Issues Arising from Reference Task Based Evaluation**

While our task-based approach has generally been successful, some issues arose in applying the method. One issue concerns our *choice of metrics* and the *importance* we associate with each. We use *multiple* evaluation metrics, in contrast to approaches, such as the PARADISE method for evaluating interactive spoken language systems (Walker et al., 1998). Our decision was influenced by several factors. Selecting appropriate evaluation metrics is a highly complex process that has generated much previous debate (Gray et al., 1993, Gray & Saltzman, 1998, Roberts & Moran, 1983, Walker et al., 1998). Prior evaluation work has, for example, shown inconsistencies between *objective measures* (time to solution and task success) and *subjective measures* (user satisfaction) for people doing the same task using the same system (Sellen, 1992, Whittaker et al., 1993). This inconsistency may make it impossible to have one metric "stand in" for another. Other evaluation work has made strong claims for the use of *user satisfaction* in evaluating system success (Walker et al., 1998), based on the argument that persistent long term system use is motivated by *user's perception* of the system's value, rather than externally calculated measures<sup>5</sup>. While acknowledging this argument, there are still outstanding questions concerning the definition and measurement of user satisfaction. Our current (conservative) view is therefore that multiple objective and subjective metrics should be used to measure system success. We regard as research questions: the exact relationship between different

---

<sup>5</sup> This is an oversimplification of Walker et al. (1998). They argue that multiple factors contribute to system success (task completion, time to solution, speech recogniser accuracy, use of help prompts), but in modelling the contribution of these factors, their regression analyses treat user satisfaction as the dependent variable. In other words, they view user satisfaction as the critical metric and their question is how these other factors affect it.

measures; whether one metric is more useful and predictive than others; how user satisfaction is defined and measured.

A second issue concerns *reference task selection*. One of our chosen tasks, summarisation, was relatively insensitive to different user interface techniques. While our requirements data revealed that summarisation was a critical task for users, summarisation performance has not proved to be a useful way to distinguish between different user interfaces for any of our metrics. Does this mean that summarisation is a poor candidate for a reference task? Closer examination of our data suggest possible reasons for our failure to find effects. Overall performance on the summarisation task was low. Our current UI techniques may not have helped with summarization, but better techniques might improve performance and produce observed differences on this task. Another possibility is that our definition of summarisation is underspecified, so the task was not well defined for users (Sparck-Jones, 1998). Our experience with summarisation has an important implication for the reference task approach. It is not enough to select important tasks by careful analysis of user data; these tasks also must be well operationalised for evaluation purposes. Operationalisation itself may be a complex undertaking to achieve plausible instantiations of tasks in experimental settings.

Another problem concerns the relationship between requirements gathering and *reference task selection*. Most requirements gathering takes place in the context of specific applications. In our case, we gathered information about speech retrieval by investigating voicemail users. But the primary function of voicemail is an asynchronous communications application rather than a speech archive. One decision when selecting reference tasks was therefore whether the observed tasks were relevant to speech retrieval or asynchronous communication. In our requirements gathering for voicemail we actually identified two further tasks, namely status tracking and archive management. We excluded these from the speech retrieval reference task set because they did not directly concern retrieval. Of course if we were identifying reference tasks for *managing asynchronous communications* (e.g. email and voicemail) then such tasks would be highly relevant.

We also experienced the problem of *task granularity*. In processing voicemail users carry out activities that are analysable at multiple levels of abstraction. At the highest level “processing voicemail” might be an activity that users engage in. Alternatively we might describe low level acts such as “press button 3” (to delete a message). Neither characterisation would be useful as a reference task. The “process voicemail” characterisation is both too general and it includes tasks that are not directly relevant to speech retrieval (status tracking and archive management). In contrast the “press button 3” characterisation is too specific to a particular implementation. In identifying our three reference tasks we made decisions about abstraction, and our criteria were intuitive. A critical technical issue for our research program concerns principled ways to specify reference task granularity.

We also should be concerned about *task-specificity*. We found different performance for search, summarisation and information extraction tasks. Different

user interface techniques may be successful for different reference tasks. Such a conclusion would indeed be consistent with observations about task-specific interfaces (Nardi, 1993), and also current theories of situated cognition (Lave, 1988, Suchman, 1987). Task-specificity highlights the importance of careful task selection. We must choose reference tasks to be critical to our users' everyday computing activities. Careful task selection ensures that we still generate important data to help improve system design for important user problems, even if that design does not generalise to *all* user tasks.

Of course, we hope our approach leads to general principles for UI design, but if not, at least we have data about important user tasks. In the worst case it might mean that the field of HCI splinters into multiple task-based research areas, but at least those areas would be informed by well researched user needs about critical user problems, along with well defined evaluation metrics. Furthermore, a number of factors would still unite such task-based communities, including methodologies such as user-centered and participatory design, modelling techniques such as GOMS, broad frameworks such as activity theory, and computational tools such as rapid prototyping environments and interface builders. And as far as application design and development is concerned, having task specific information may correspond well with common practice: application development takes place in a task specific context.

Another issue concerns user population. While we have made every attempt to ensure the representativeness of the people participating in our experiments, specific user groups (e.g. elderly people or children) may use the technology quite differently. User population also must be included in the reference task analysis.

Another issue concerns inherent limitations of task-based evaluation. In experimental studies people are asked to perform pre-specified tasks over a short period of time. We therefore cannot detect ad-hoc or creative usage of the UI, nor how usage strategies evolve over time. These phenomena can only be observed in field trials. Of course field trials also have their drawbacks. Field trial users select their own tasks making it impossible to draw direct comparisons between different techniques or systems because users are executing different tasks. Extended usage in field trials should therefore be used to complement task based evaluation. The entire evaluation process must be iterative, combining the results of experimental and field based methods. Field trials may show that critical user tasks have been neglected, or that technologies may be developed and used in novel ways. Field trial findings should therefore be used to modify existing task definitions for future evaluations.

Finally, there is a question of novelty. What's new about the reference task speech browsing and retrieval example? After all, isn't the process we just described, good, but standard, HCI practice? Isn't it standard best practice in HCI to interview users to understand their needs, develop a system to meet these needs, and evaluate the system to determine whether it meets their needs? It seems however that there are major differences between ideal and actual descriptions of the process of HCI. Although the ideal is to follow the three steps we describe, few actual studies seem to execute all three. Recall also, that the reference task

agenda involves both technical and social aspects. We diverge from standard technical practice in recommending that we use general evaluation metrics, derived from important tasks. However, the more important implications of our example are social. In our domain, we found no set of task definitions or user requirements, and no accepted metrics. And in moving toward developing this knowledge, there were no accepted community mechanisms for refining and disseminating that knowledge once we had discovered it. Developing such social mechanisms is the major activity needed to put the reference task approach into practice.

## 6. CONCLUSIONS

We identify a problem with the process of HCI research: emphasis on radical innovation precludes building a common research focus. Without such a focus, people cannot build on the work of others, or compare UI techniques, in order to improve them. The lack of common focus also makes it difficult to accumulate the research on common problems needed for theory development. Lack of common knowledge also means that we cannot give informed design advice to builders of new systems. In response to this, we argue that the HCI community should focus around reference tasks. We review the advantages and disadvantages of this approach, documenting its use in information retrieval and speech recognition research. We also describe an example reference task for searching and browsing speech archives. We point to a number of outstanding issues arising from applying the approach: choice of metrics, selection and operationalisation of tasks, task-specificity of results, user variability and the need for complementary field trials. We also point to the absence of methods for distributing and sharing data and results within the field.

We also outline the necessary steps to execute the reference task research agenda. We make both technical and social recommendations. The necessary technical research involves: identifying important user tasks by systematic requirements gathering; definition and operationalisation of reference tasks and evaluation metrics; execution of task-based evaluation along with judicious use of field trials. The major technical hurdles are: (a) agreeing on common task definitions; (b) developing general templates for describing reference tasks, stating the criteria they must satisfy, including their level of granularity; (c) defining appropriate metrics; (d) designing appropriate task-based evaluation techniques.

Perhaps more important, we recommend changes in HCI community practice. We must create influential forums for discussion of common tasks and methods by which people can compare systems and techniques. The major obstacle here is to define a process that (a) allows researchers to agree on task definitions and (b) provide methods to disseminate these definitions so that they are broadly used by the HCI community. Only by doing this can reference tasks be incorporated into the process of research and development, helping the field achieve the focus it desperately needs.

## 7. ACKNOWLEDGEMENTS

Thanks to Julia Hirschberg, Candy Kamm, Fernando Pereira and Marilyn Walker, along with the attendees at HCIC 1999 for useful suggestions and feedback.

## 8. REFERENCES

- Amento, B., Hill, W., Terveen, L., Hix, D., & Ju, P. (1999). An empirical evaluation of user interfaces for topic management of Web sites. In *Proceedings of CHI '99 Conference on Computer Human Interaction*, 552-559, New York: ACM.
- Baecker, R. (1987). Towards an effective characterization of graphical interaction. In R. Baecker & W. Buxton (Eds.), *Readings in Human Computer Interaction*, San Francisco, CA: Kaufmann.
- Barreau, D. & Nardi, B. (1995). Finding and reminding: organization of information from the desktop, *SIGCHI Bulletin*, 27, 39-45.
- Bickhard, M.H. & Terveen, L.G. (1995). *Foundational Issues in Artificial Intelligence and Cognitive Science: Impasse and Solution*. New York: Elsevier.
- Brennan, S. (1990). Conversation as direct manipulation: an iconoclastic view. In B. Laurel (Ed.), *The art of human computer interface design*. Reading, MA: Addison-Wesley.
- Burkhart, B., Hemphill, D., & Jones, S. (1994). The value of a baseline in determining design success. In *Proceedings of CHI '94 Conference on Computer Human Interaction*, 386-391, New York: ACM.
- Carroll, J. & Campbell, R. (1986). Softening up hard science. *Human Computer Interaction*, 2, 227-249.
- Chapanis, A. (1975). Interactive human communication. *Scientific American*, 232, 36-42.
- Choi, J., Hindle, D., Hirschberg, J., Magrin-Chagnolleau, I., Nakatani, C. H., Pereira, F., Singhal, A., & Whittaker, S. (1998). SCAN - Speech content audio navigator: a systems overview. In *Proceedings of International Conference on Spoken Language Processing*, 604-608, Piscataway, NJ: IEEE.
- Communications of the ACM*, Special Issue on Recommender Systems, 40, Resnick, P., & Varian, H.R., guest editors.
- Dreyfus, H.L. (1992). *What Computers Still Can't Do*. Cambridge, MA: MIT Press.
- Fertig, S., Freeman, E., & Gelertner, D. (1996). Finding and reminding reconsidered. *SIGCHI Bulletin*, 28 (1).
- Ford, K.M. & Pylyshyn, Z. (1995). *The Robot's Dilemma Revisited: The Frame Problem in Artificial Intelligence*. Norwood, NJ: Ablex Press.
- Foundyleer, C. (1984). *Cad/CAM, CAE: The Contemporary Technology*. Cambridge, MA: Daratech Associations.

- Gifford, D., Jouvelot, P., Sheldon, M. & O'Toole, J. (1991). Semantic file systems. In *Proceedings of 13<sup>th</sup> ACM Symposium on Operating System Principles*, NY: ACM.
- Goldberg, D., Nichols, D., Oki, B.M. & Terry, D. (1992). Using Collaborative Filtering to Weave an Information Tapestry. *Communications of the ACM*, 35, 51-60.
- Gray W. & Salzman, M. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human Computer Interaction*, 13, 203-262.
- Gray, W. D., John, B. E., & Atwood, M. E. (1993). Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world task performance. *Human-Computer Interaction*, 8, 237-309.
- Grudin, J. (1988). Why CSCW applications fail: problems in the design and evaluation of organizational interfaces. In *Proceedings of CSCW' 88 Conference on Computer Supported Cooperative Work*, 85-93, New York: ACM .
- Harnad, S. (1990) The Symbol Grounding Problem. *Physica D* 42, 335-346.
- Hill, W.C., Stead, L., Rosenstein, M. & Furnas, G. Recommending and Evaluating Choices in a Virtual Community of Use, in *Proceedings of CHI'95 Conference on Computer Human Interaction*, 194-201, New York, ACM.
- Isaacs, E. & Tang, J. (1996). Technology transfer: so much research so few good products. *Communications of the ACM*, 39, 22-25.
- Junqua, J-C. (1999). The Lombard Effect: A reflex to better communicate with others in noise. *International Conference on Acoustics Speech and Signal Processing*, 2083-2086, Piscataway, NJ: IEEE.
- Kaptelinin, V. (1996). Activity Theory: Implications for Human-Computer Interaction. In B. Nardi, (Ed.), *Context and Consciousness: Activity Theory and Human-Computer Interaction*. Cambridge, Mass.: MIT Press.
- Kraut, R., Fish, R., Root, B. & Chalfonte, B. (1993). Informal communication in organizations. In R. Baecker (Ed.), *Groupware and Computer Supported Cooperative Work*. San Francisco, CA: Kaufman.
- Kuhn, T. S. (1996). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
- Landauer, T. (1995). Let's get real. In R. Baecker, J. Grudin, W. Buxton & S. Greenberg (Eds.), *Human Computer Interaction: towards the year 2000*. San Francisco, CA: Morgan Kaufman.
- Laurel, B. (1990). *The art of human computer interface design*. Reading, MA: Addison-Wesley.
- Lave, J. (1988). *Cognition in practice*. New York: Cambridge University Press.
- Marcus, M. (1991) *Proceedings of Speech and Natural Language Workshop*, San Francisco, Kaufmann,
- Marvin, C. (1988). *When old technologies were new*. New York: Oxford University Press.

- Nakatani, C. H., Whittaker, S., & Hirschberg, J. (1998). Now you hear it now you don't: empirical studies of audio browsing behavior, In *Proceedings of International Conference on Spoken Language Processing*, 1003-1007, Piscataway, NJ: IEEE.
- Nardi, B. (1993). *A small matter of programming*. Cambridge, MA. : MIT Press.
- Nardi, B., Kuchinsky, A., Whittaker, S., Leichner, R. & Schwarz, H. (1996). Video-as-data: Technical and social aspects of a collaborative multimedia application. *Computer Supported Cooperative Work*, 4, 73-100.
- Nardi, B., Miller, J. & Wright, D. (1998). Collaborative, Programmable Intelligent Agents. *Communications of the ACM*.
- Newell, A & Card, S. (1985). The prospects for psychological science in Human Computer Interaction, *Human Computer Interaction*, 1, 209-242.
- Newman, W. (1994). A preliminary analysis of the products of HCI research using Pro Forma abstracts, In *Proceedings of CHI '94 Conference on Computer Human Interaction*, 278-284, New York: ACM.
- Newman, W. (1997). Better or just different? On the benefits of designing interactive systems in terms of critical parameters. In *Designing Interactive Systems (DIS97)*, 239-246, New York: ACM.
- Olson, J., & Olson, G. (1990). The growth of cognitive modeling in human computer interaction since GOMS. *Human Computer Interaction*, 5, 221-265.
- Orlikowski, W. (1992). Learning from Notes: Organizational issues in groupware implementation. In *Proceedings of CSCW '92 Conference on Computer Supported Cooperative Work*, New York: ACM.
- Oviatt, S.L., Levow, G., MacEachern, M., & Kuhn, K. (1996). Modeling hyperarticulate speech during human-computer error resolution. In *Proceedings of the International Conference on Spoken Language Processing*, 801-804. Piscataway, NJ: IEEE.
- Price, P. (1991) *Proceedings of Speech and Natural Language Workshop*, San Francisco, CA., Kaufmann.
- Rao, R., Card, S., Johnson, W., Klotz, L., & Trigg, R. (1994). Protofoil: Storing and finding the information worker's documents in an electronic filing cabinet. In *Proceedings of CHI '94 Conference on Computer Human Interaction*, 180-185, New York: ACM.
- Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J. (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of CSCW '94 Conference on Computer Supported Cooperative Work*, 175-186. New York: ACM.
- Roberts, T. L., & Moran, T. P. (1983). The evaluation of text editors: Methodology and empirical results. *Communications of the ACM*, 26, 265-283.
- Rudisill M., Lewis C. L., Polson P. G. & McKay T. D. (1996). *Human-Computer Interface Design: Success Stories, Emerging Methods and Real-World Context*. San Francisco: Kaufmann.

- Searle, J.R. (1981). Minds, Brains, and Programs. In J. Haugeland (Ed.), *Mind Design*. 282-306, Cambridge: MIT.
- Sellen, A. (1995). Remote conversations: the effects of mediating talk with technology, *Human Computer Interaction*, 10, 401- 444.
- Shneiderman, B. (1982). The future of interactive systems and the emergence of direct manipulation, *Behavior and information technology*, 1, 237-256.
- Shardanand, U., & Maes, P. (1995). Social Information Filtering: Algorithms for Automating “Word of Mouth”. In *Proceedings of CHI’95 Conference on Computer Human Interaction*, 210-217, New York: ACM.
- Smith, D., Irby, C., Kimball, R., Verplank, W., & Harslem, E. (1982). Designing the Star user interface. *Byte*, 7.
- Sparck-Jones, K. (1998). Summary performance comparisons TREC2, TREC3, TREC4, TREC5, TREC 6. In Voorhees, E. M., & Harman, D. K. (Eds.), *Proceedings of the Sixth Text Retrieval Conference (TREC-7)*, 1998.
- Spark-Jones, K. (1998). Automatically summarising: factors and directions. In I. Mani & M. Maybury (Eds.), *Advances in Automatic Text Summarization*. Cambridge, MA., MIT Press.
- Stark, L., Whittaker, S, and Hirschberg, J. (2000). ASR satisficing: the effects of ASR accuracy on speech retrieval. To appear in ICSLP 2000.
- Stern, R. (1990). *Proceedings of Speech and Natural Language Workshop*, San Francisco, CA: Kaufmann.
- Suchman, L. (1987). *Plans and situated actions*. Cambridge University Press, Cambridge.
- Sutherland, I. (1963). Sketchpad: A Man-Machine Graphical Communication System. *Proceedings of AFIPS 23*, 329-346.
- Voorhees, E. M., & Harman, D. K., (1998). Overview of the seventh Text Retrieval Conference (TREC-7), in Voorhees, E. M., & Harman, D. K. (Eds.), *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*, 1998.
- Voorhees, E. M., & Harman, D. K., (1997). Overview of the sixth Text Retrieval Conference (TREC-6), in Voorhees, E. M., & Harman, D. K. (eds.), *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*, 1997.
- Walker, M., Litman, D., Kamm, C., & Abella, A. (1998). Evaluating Spoken Dialogue Agents with {PARADISE}: Two Case Studies. *Computer Speech and Language*, 12, 3.
- Wayne, C. (1989). *Proceedings of Speech and Natural Language Workshop*, San Francisco, CA: Kaufmann.
- Whittaker, S., Choi, J., Hirschberg, J., & Nakatani, C. (1998a). What you see is almost what you get: design principles for user interfaces for speech archives. In *Proceedings of the International Conference on Speech and Language Processing*, 1009-1013, Piscataway, NJ: IEEE.

Whittaker, S., Davis, R., Hirschberg, J., and Muller, U. (2000). Jotmail: a voicemail interface that enables you to see what was said. In *Proceedings of CHI'2000 Human Factors in Computing Systems*, 89-96, New York: ACM.

Whittaker, S., Frohlich., & Daly-Jones, O. (1994). Informal workplace communication: what is it like and how might we support it? In *Proceedings of CHI'94 Human Factors in Computing Systems*, 130-137, New York: ACM.

Whittaker, S., Geelhoed, E., & Robinson, E. (1993). Shared workspaces: how do they work and when are they useful? *International Journal of Man-Machine Studies*, 39, 813-842.

Whittaker, S., Hirschberg, J., Choi, J., Hindel, D., Pereira, F., & Singhal, A. (1999). SCAN: designing and evaluating user interfaces to support retrieval from speech archives. In *Proceedings of SIGIR99 Conference on Research and Development in Information Retrieval*, 26-33, New York: ACM.

Whittaker, S., Hirschberg, J., & Nakatani, C.H. (1998b). All talk and all action: strategies for managing voicemail messages. In *Proceedings of CHI '98 Conference on Computer Human Interaction*, 249-250, New York: ACM .

Whittaker, S., Hirschberg, J., & Nakatani, C. H. (1998c). What you see is almost what you hear: design principles for user interfaces for accessing speech archives, In *Proceedings of International Conference on Spoken Language Processing*. Piscataway, NJ: IEEE.

## **List of figure titles**

*Figure 1: Mean average precision of different Cornell systems for lifetime of TREC*

*Figure 2 – WYSIAWYH browser providing Overview and Transcript for browsing*

*Figure 3 – Basic browser providing play controls for browsing*

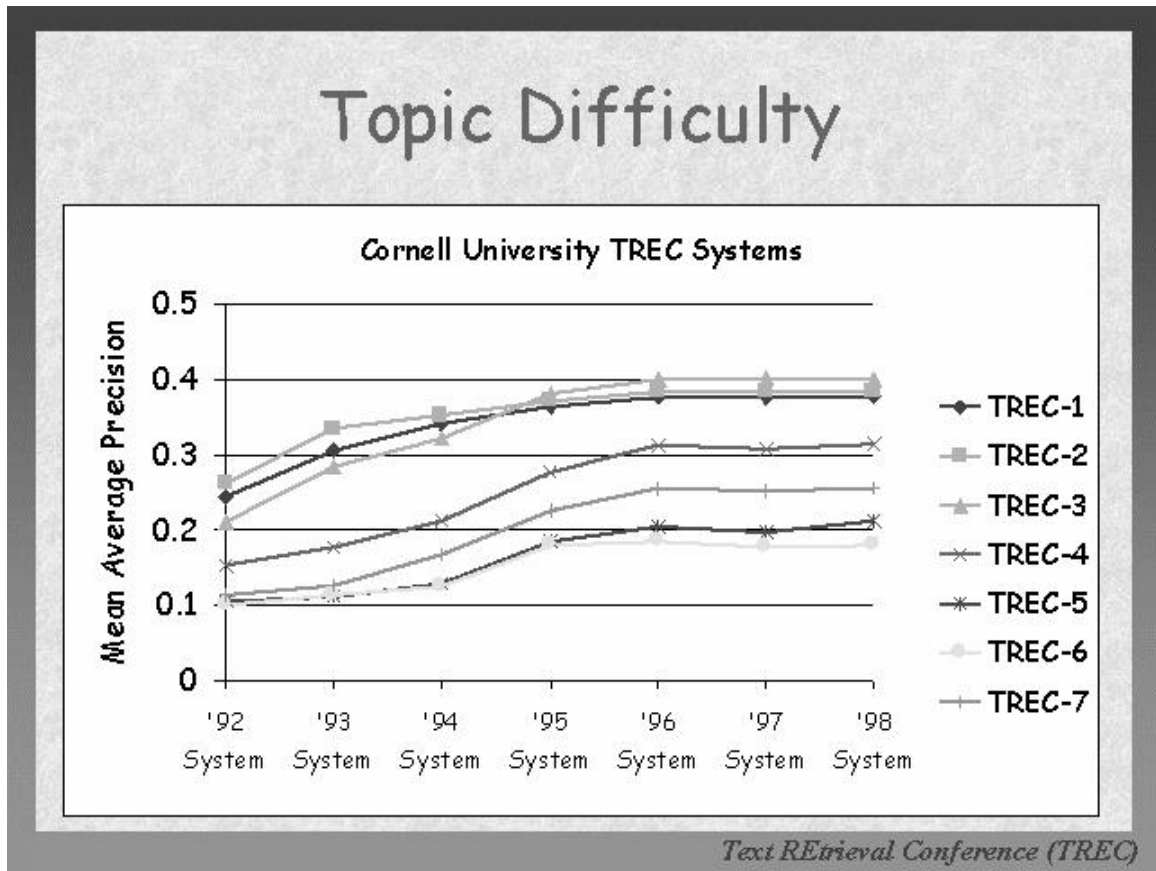


Figure 1: Mean average precision of different Cornell systems for lifetime of TREC

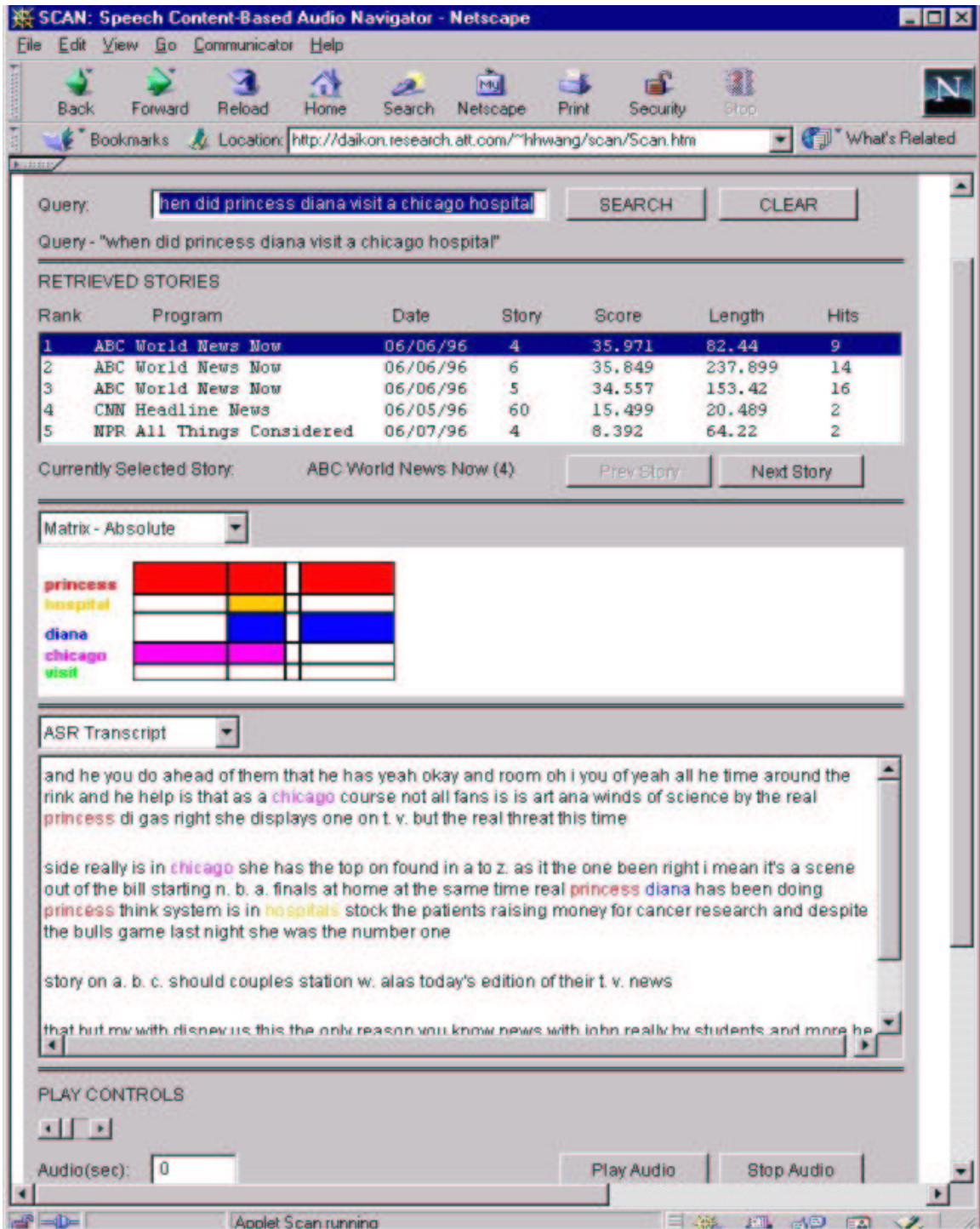


Figure 2 – WYSIAWYH browser providing Overview and Transcript for browsing

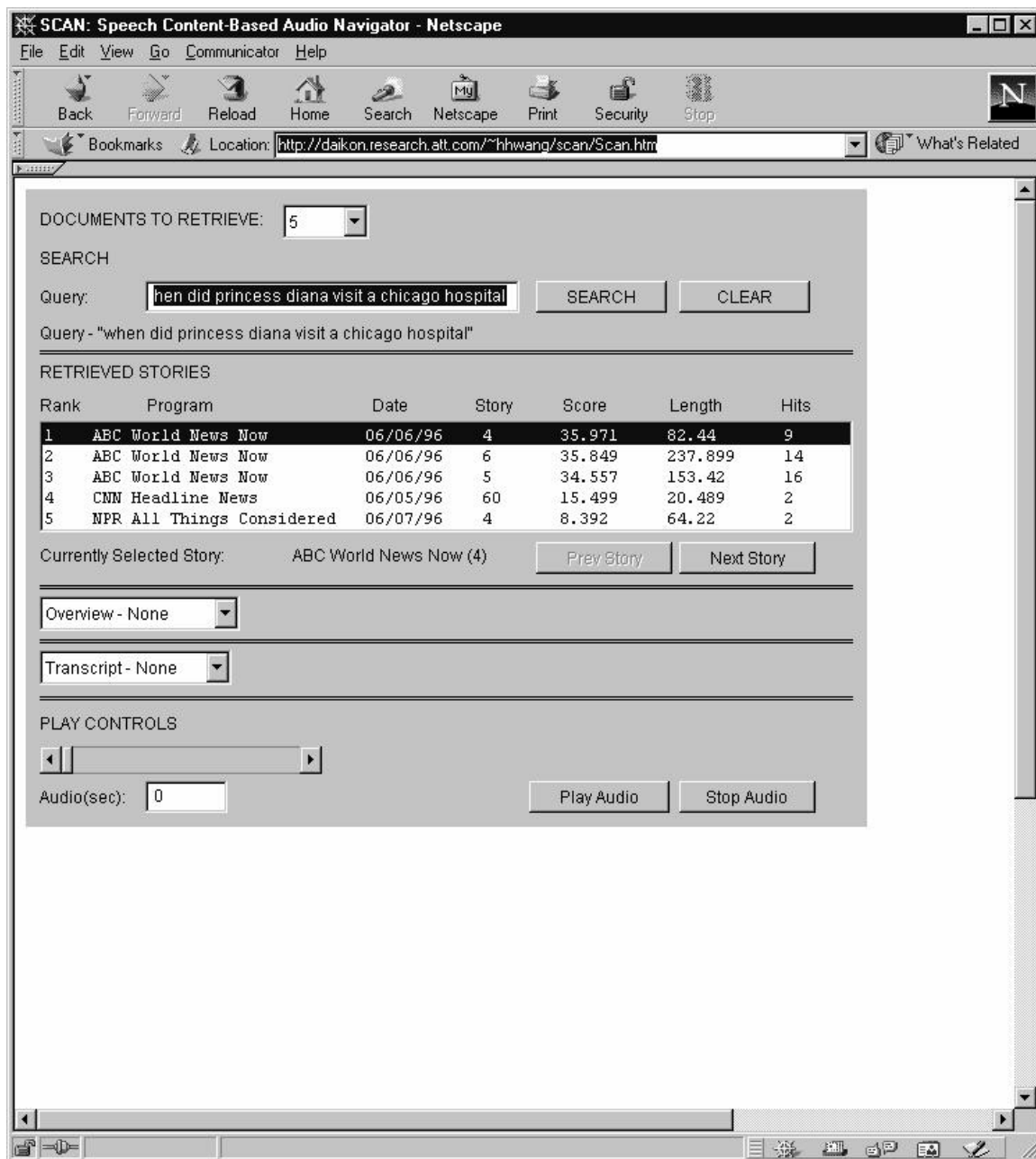


Figure 3 – Basic browser providing play controls for browsing

