WHAT YOU SEE IS (ALMOST) WHAT YOU HEAR: DESIGN PRINCIPLES FOR USER INTERFACES FOR ACCESSING SPEECH ARCHIVES

Steve Whittaker, John Choi, Julia Hirschberg, Christine H. Nakatani

ATT Labs-Research, 180 Park Ave., Florham Park, NJ, 07932, USA
email: {stevew,choi,julia,chn@research.att.com}

ABSTRACT

Despite the recent growth and potential utility of speech archives, we currently lack tools for effective archival access. Previous research on search of *textual* archives has assumed that the system goal should be to retrieve sets of relevant documents, leaving users to visually scan through those documents to identify relevant information. However, in previous work we show that in accessing real speech archives, it is insufficient to only retrieve "document" sets [9,10]. Users experience huge problems of *local navigation* in attempting to extract relevant information from within speech "documents". These studies also show that users address these problems by taking handwritten notes. These notes detail both the *content* of the speech and serve as *indices* to help access relevant regions of the archive. From these studies we derive a new principle for the design of speech access systems: What You See Is (Almost) What You Hear. We present a new user interface to a broadcast news archive, designed on that principle.

1. INTRODUCTION

Recently there have been huge increases in the amounts of personal and public data stored in digital speech archives. Broadcasting companies such as PBS and the BBC have made radio programs available, and various types of public records (e.g. US Congressional Debates) are also being archived. Furthermore, research indicates that such archives are potentially highly valuable: speech communication has been shown to be both ubiquitous and critical for the execution of many workplace tasks [6]. However, despite the potential importance of such speech archives, they are currently underutilized, due to the lack of useful tools for accessing and browsing large speech archives.

One natural starting point for identifying techniques for speech access is in the information retrieval literature, which represents over 20 years of research into the retrieval of text documents from large corpora [8]. Yet with few exceptions [5], the focus of textual information retrieval techniques has been on *search*, with the goal of identifying *sets of documents* that are relevant to a given user query. Current information retrieval techniques do not support other types of information seeking behaviors that require users to find local information, e.g. extracting specific facts or identifying regions of interest from *within a document*. Consistent with this focus on search, user interfaces to information retrieval systems typically present a relevance ranked set of documents in response to a user query. They assume that for more detailed information seeking, users can easily visually scan and browse through textual documents, for example, when they need to identify specific regions of interest within a document.

In the context of a *speech* corpus, however, it is clear that a user interface supporting only search is insufficient, given the complexity of *local navigation* within speech stories. The news stories in our corpus can be as long as 15 minutes. Given the sequential nature of speech, it is difficult to quickly scan through long speech stories to obtain an overview of the contents of a story [1], or to identify specific information of direct relevance [3]. It is therefore both inefficient and inappropriate to expect users to listen to multiple lengthy stories in their entirety, when relevant information may be located in a specific portion of a particular story. This indicates that, in addition to search, interfaces for accessing speech archives need to support local navigation specifically: story scanning and information extraction. This paper describes two user studies of voicemail which (a) examine the problems of local navigation; and (b) identify the strategies users employ to overcome these problems. From these studies we derive a new principle for the design of speech access systems: What You See Is (Almost) What You Hear. We present a new user interface to a broadcast news archive, designed on that principle.

2. STUDIES OF LOCAL NAVIGATION: SCANNING AND INFORMATION EXTRACTION STRATEGIES

To better understand user requirements for local navigation, we conducted two studies of speech browsing by studying voicemail access. One critical motivation was to study *realistic* access behaviors. Voicemail represents a domain with experienced users who have evolved strategies for dealing with important speech data. Voicemail therefore represents a good starting place for identifying user problems with local navigation in speech.

In one experiment we examined local navigation strategies under controlled laboratory conditions. We gave users two different types of simple graphical user interfaces to a voicemail archive [10]. The first was based on a tape-recorder metaphor, in which the speech archive was represented as a horizontal bar. If users wanted to access a given point in the audio, (e.g halfway through) they could place a cursor at that point in the bar and press the play button. The second UI was similar except that it had CD type "tracks" added, to indicate the beginning of each new message. In both cases the motivation was to design simple interfaces around familiar metaphors. Users were given two types of access tasks, based on interviews and surveys conducted with voicemail users

[9]. In one task they had to summarize a voicemail message, and in the other they had to extract information (e.g. a fact such as a name, date or phone number) from a message. We found that users experienced huge problems with local navigation, even for a small archive of eight messages where the average message length was around 30s. Specifically, we found that: (a) providing structural information was less helpful than predicted; (b) users did not seem able to learn the contents of a given message - they were no faster to answer questions about familiar than novel messages; (c) information extraction tasks were extremely hard - when answering questions requiring access to two facts within a given message users repeatedly replayed material they had just heard, suggesting they had forgotten what they had just played; (d) in a post-hoc memory task, users showed poor recall for the contents of the messages. These findings underscore the problems users experience with local navigation, even in a small archive.

A second study [9] used a combination of interview and survey methods to investigate the strategies that people employ to access information from voicemail archives. We investigated 148 high volume users (defined as those receiving more than 10 messages per day). Note-taking was a central voicemail processing strategy and 72% of users said that they 'almost always' take notes. Users employed two different note-taking strategies. The first strategy was *full transcription*: here users attempt to produce a verbatim written transcript of the target message, in order to preclude future access of that message. The second strategy was to take notes as *indices*. The objective of this strategy was to abstract the key points of the message (such as caller name, caller number, reason for calling, important dates/times and action items). In most cases, users kept the original message as a backup for these incomplete and sometimes sketchy notes. Others routinely used the *indexing* strategy and kept their handwritten notes sequentially in a note-pad. They referred to this index list, when searching their archive to locate particular messages. Without such an index, they would have had to re-access each message and listen to it in detail. Finally we examined the different cues people used for processing voicemail. We discovered the importance of *intonation.* Intonation serves as: (a) an indicator of urgency, with people stating that they could tell the importance of a given message within a few seconds of listening to it; (b) a way of clarifying speaker intentions to determine what a person really meant.

What are the implications of these two studies for the design of interfaces to spoken news archives? The studies underscore users' problems with local navigation. These problems emerged even with voicemail messages that are characteristically short (the maximum length of a message on the system we studied was two minutes). The note-taking strategies also suggest techniques for addressing these problems. *Indexing* provides an abstract overview of each message, in terms of its key points. *Indexing* also serves as a guide when navigating the archive. The *full-transcription* strategy suggests that having a textual rendering of speech can provide a rapid way to extract information from a given message. Our user comments also suggest, however, that people do not want to be completely reliant on such a *transcript:* given the importance of intonation, people want to still be able to refer to the original message.

3. THE SCAN USER INTERFACE TO THE NEWS STORY RETRIEVAL SYSTEM

As we have noted, the information retrieval literature underscores the importance of supporting *search,* but our own experiments on speech archives indicate the critical role of scanning and information extraction for speech information. The SCAN (system for content-based audio navigation) user interface therefore has three components: *search, overview* and *transcript,* with overview and transcript elements providing support for local navigation. A screen dump of the user interface is shown in Fig 1. The underlying system supports access to a broadcast news corpus. For further information see [2]. We now describe each element of the UI in detail.

3.1 Search

The function of search in SCAN is to provide rapid access to a set of potentially relevant transcripts and the original speech stories corresponding to each. How is search accomplished? The backend system takes the speech stories corpus, and applies some preprocessing (including the segmentation of each story into 'audio paragraphs' on the basis of acoustic-prosodic information). We then apply ASR to each story. The result is a set of errorful transcripts. Each transcript is next indexed using the SMART retrieval engine. When the user executes a query against the indexed transcript corpus, SMART returns list of stories ranked in order of their relevance. For a full system description, the reader should consult [2].

The search panel is at the top of the browser. In order to search for news programs about a given topic, the user types their query into the query window and presses "search" in order to execute the query. In the Figure the user has typed in the query "What is the status of the trade deficit with Japan?". The results are depicted in the results panel immediately below the search panel. The results panel presents a list of 10 news stories ranked in order of their relevance to the query. There is also additional information about each story: the program name, the program number (there are often multiple programs from the same station on the same day), the date, the relevance rank score, the length (in seconds), and the total number of instances of the query words ("hits") that appear in the story. The user can select a story of interest by clicking on it. The text highlighting shows the currently selected story. Users can also move through the list of stories using the "previous doc" and "next doc" buttons.

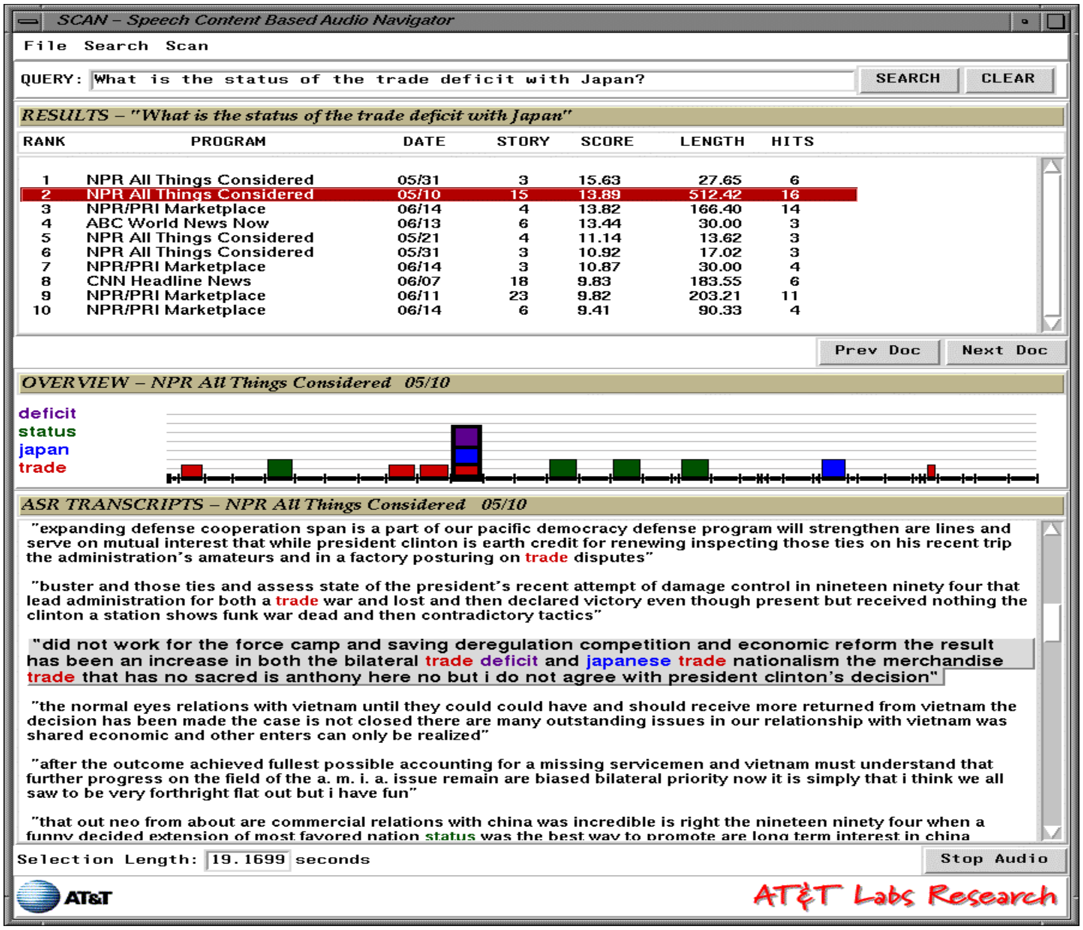

Figure 1: The SCAN user interface

3.2 Overview

The function of the overview is to provide high level information about each story. Overview data allows users to rapidly scan within stories to identify regions that are of particular relevance. Overview information is presented in the form of the query word distribution within each story. This distribution is shown in the overview panel, which highlights regions where query words are most prevalent. Each query word is color coded, and each segment ('audio paragraph') in the story is represented by a vertical column in a histogram. The width of the bar represents the relative length of that story segment. The height of each histogram bar represents the overall query word density within that segment (number of instances of the query words normalized for the segment length). Users can also determine the distribution of specific query words (e.g. "japan" as represented by the dark tile, occurs only in segments 11 and 27). A similar technique is used for textual documents in [5]. The user can access the speech for any of these segments simply by clicking on the bar representing the segment. Selecting a bar initiates play from the start of the relevant segment.

3.3 Transcript

The function of the transcript is to support information extraction, by providing access to detailed information about what was said in the story. The transcript panel shows a transcription of the selected story "NPR: All things considered". Because the transcript has been generated using automatic speech recognition technology, it contains errors (e.g. in the first paragraph the ASR misrecognized the word "Japan" as "span", and in paragraph 4 " the normal eyes" should read "to normalize"). Query words in the transcript are highlighted and color coded, and it is possible to play a given segment of the story by clicking on the relevant paragraph in the transcript. The paragraph is highlighted as it plays. Playback can be stopped at any point either by selecting a different paragraph, or by pressing the "stop audio" button.

The transcript has several functions. First, if it is accurate, users may be able to find the information they need simply by reading the transcript without listening to the audio. Other functions of the transcript are that like the overview, it provides a method for rapidly scanning the speech to find regions of relevance: users can visually skim through the transcript to find areas of interest. Unlike the overview, however, the transcript also provides *local contextual* information: users can play a given 'audio paragraph' and simultaneously read the text for surrounding paragraphs to determine local context. Finally, reading the transcript can allow users to make judgments about the accuracy of search and overview information. If the transcript contains bizarre word combinations and grammar (e.g. paragraph 2), this suggests that query words may also have been misrecognized, so that users should be less trustful of search and overview information provided elsewhere in the system. If errors are prevalent then users should rely more on the speech rather than the transcripts.

Together the elements of the user interface support a new design principle for speech retrieval interfaces: "what you see is (almost) what you hear" (WYSIAWYH). A key element of the user interface lies in providing a *visual analogue* to the underlying speech. By depicting the speech as text in the form of the overview and transcript we support *visual* scanning and information extraction. Providing this v*isual* information therefore addresses some of the problems of local navigation within speech data that we identified in our user studies. But could we not rely on a purely textual interface without needing access to speech? There are two reasons why a purely textual representation might be insufficient: transcription errors and the importance of intonation. The existence of ASR errors means that the transcript frequently diverges from the underlying speech. This discrepancy explains the use of the term "almost" in the WYSIAWYH principle: the speech users hear may not correspond with the transcript provided. We had around 30% ASR word recognition errors for this corpus, and given the intractability of the Out of Vocabulary problem, such errors will persist, even with improvements in recognition technology. Users also stressed the importance of *intonation* and the need to *hear exactly how* something was said. Recognition errors and the importance of intonation together mean that text alone is insufficient and we must always offer users access to underlying speech.

4. RELATED WORK AND FUTURE RESEARCH

A number of speech search systems have recently been built to compete in the TREC forum [8]. However, the main focus of these systems has generally been within the information retrieval tradition of developing methods for efficiently finding sets of speech documents, rather than supporting local navigation tasks. Other systems have also been built to provide access to multimedia data such as television broadcasts [4,6]. These systems use vision techniques such as scene analysis and key frame detection to identify transitions and scenes in video. Users can click on salient video stills to access the underlying speech relating to a given scene. Such systems have sometimes relied on close captioning rather than ASR for speech search. One exception is the Infomedia system [4], which has used speech skimming [1] in supporting local navigation on ASR generated speech.

Elsewhere we have reported our results concerning the effectiveness of speech *search* using traditional information retrieval metrics and tasks [8]. We are currently conducting experiments to evaluate the interface, particularly how well it supports *local navigation*. We are testing the user interface on three different information access tasks which are intended to test different aspects of local navigation: (a) selecting which of five speech stories is the most *relevant* to a given topic; (b) *summarizing* a given story; (c) *finding a fact* from a given story. The first task is intended to test user's ability to scan whole stories; the second task to test both scanning and information extraction; and the final task to test information extraction. We will compare the SCAN browser with a basic browser offering *search,* but little support for local navigation.

5. REFERENCES

1. Arons, B. *Interactively skimming speech*. Unpublished PhD thesis, MIT Media Lab, 1994.
2. Choi, J., Hindle, D., Hirschberg, J., Magrin-Chagnolleau, I., Nakatani, C., Pereira, F., Singhal, A., Whittaker, S., An Overview of the AT&T Spoken Document Retrieval System, *Proceedings of the DAARPA/NIST Broadcast News Transcription and Understanding Workshop*, 1998, (forthcoming).
3. Gould, J. Human factors challenges – The speech filing system approach. In *ACM Transactions on Office Information Systems*, 1(4), October 1983.
4. Hauptmann, A. and Witbrock, M. Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval. In Maybury, M., ed. *Intelligent Multimedia Information Retrieval*, AAAI Press, 1997.
5. Hearst, M. Tilebars: Visualization of term distribution in full text information access. In *Proceedings of CHI'95 Human Factors in Computing Systems*, ACM Press, New York, 1995.
6. Kraut, R., Fish, R., Root, B., and Chalfonte, B. Informal communication in organizations. In R. Baecker (Ed.), *Groupware and Computer Supported Co-operative Work*, 287-314, Morgan Kaufman, 1992.
7. Shahraray, B., and Gibbon, D. C. Automated authoring of hypermedia documents of video programs. In *Proceedings of the Third ACM Conference on Multimedia*, 401-409, San Francisco, 1995.
8. Voorhees, E. M., and Harman, D. K., Overview of the sixth Text Retrieval Conference (TREC-6), in Voorhees, E. M., and Harman, D. K. (eds.), *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*, 1998, forthcoming.
9. Whittaker, S., Hirschberg, J. & Nakatani, C. All talk and all action: Strategies for managing voicemail data. In *Proceedings of CHI'98 Human Factors in Computing Systems*, ACM Press, New York, 1998.
10. Whittaker, S., Hirschberg, J. & Nakatani, C. Play it again: a study of the factors underlying speech browsing behaviour. In *Proceedings of CHI'98 Human Factors in Computing Systems*, ACM Press, New York, 1998.