"I just played that a minute ago!:" Designing User Interfaces for Audio Navigation

Julia Hirschberg, John Choi, Christine Nakatani, and Steve Whittaker
AT&T Labs – Research
Florham Park NJ 07932
(julia,choi,chn,stevew)@research.att.com

The current popularity of multimodal information retrieval research critically assumes that consumers will be found for the multimodal information thus retrieved and that interfaces can be designed that will allow users to search and browse multimodal information effectively. While there has been considerable effort given to developing the basic technologies needed for information retrieval from audio, video and text domains, basic research on how people browse and search in any of these domains, let alone in some combination, has lagged behind. In developing the SCAN (Spoken Content-based Audio Navigation) system to retrieve information from an audio domain, we have attempted to study the problems of how users navigate audio databases, hand in hand with the development of the speech and information retrieval technologies which enable this navigation.[1]

SCAN was developed initially for the TREC-6 Spoken Document Retrieval (SDR) task, which employs the NIST/DARPA HUB4 Broadcast News corpus. However, we are also developing a search and browsing system for voicemail access, over the telephone and via a GUI interface. To this end, we have built several user interfaces to both the voicemail and news domains, which we are employing in a series of laboratory experiments designed to identify limiting and enabling features of audio search and browsing interfaces. We want to examine the following questions: a) how do people want to search audio data? what sort of search and play capabilities do they make most use of, when given several alternatives? b) do people search different sorts of audio data (e.g., familiar versus unfamiliar) differently? c) do people perform different types of audio search task (e.g. finding a single fact vs. summarizing a longer audio document, or finding an audio) differently? d) what are the major barriers to efficiency of audio search? what additional aids might help to overcome these? e) what design principles underly the creation of effective interfaces to audio databases?

In this paper we present a brief overview of the SCAN system, describe several browsing prototypes and the different aspects of audio browsing/retrieval they have been designed to test, and present results of two sets of experiments involving their use. We then describe two novel browsers, developed from the results of these experiments, which employ document segmentation information and errorful automatic speech recognition transcription as aids to audio browsing, and briefly outline additional experimental work on their use.

1 The SCAN System

SCAN was developed for the TREC-96 SDR task, a known item information retrieval (IR) task from approximately 47 hours of the NIST/DARPA HUB4 Broadcast News/SDR speech corpus. Like most systems participating in this task, SCAN uses automatic speech recognition (ASR) techniques to produce an (errorful) transcription of the speech and then applies text-based IR techniques on the transcription to rank the corresponding speech documents as to their relevance to a given text query. Results of the IR ranking are returned to the user via one of several interfaces, speech or text-transcription driven, which are described below.

The system architecture is shown in Figure 1. Speech documents, labeled by hand in the SDR

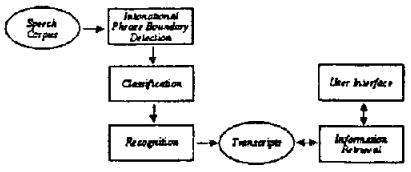

Figure 1: Architecture of the SCAN audio browsing/retrieval system.

data, are first run through an intonational phrase

[1]The SCAN audio browsing and retrieval system has been under development since June 1997 at AT&T Labs – Research, and represents collaborative work by Don Hindle, Ivan Magrin-Chagnolleau, Fernando Pereira, Amit Singhal, and the authors, with much additional help from Andrej Ljolje, Aaron Rosenberg and S. Parthasarathy.

detection procedure, to segment stories into principled "chunks" of speech. This initial segmentation provides the ASR engine with manageable units of speech to recognize, and later provides users with manageable units of speech to listen to. Each phrase identified is then classified as to its channel conditions, so that the most appropriate acoustic models can be applied during recognition; this classification is done by assigning log-likelihood scores based upon a partition of the data into 'high', 'medium', and 'low' fidelity speech (corresponding to 'studio 8K speech', '8K speech recorded under other conditions', and '4K telephone speech' – all recorded without background noise or music) and noise (all speech recorded with noise or music). The recognizer employs a time-synchronous beam search with continuous density, three-state, left-to-right, context-dependent Hidden Markov phone models. The production of word sequences from phones is implemented using weighted finite-state transducers. Recognition hypotheses are output as word lattices. Several language models have been developed from the 116 million word SDR corpus, including standard backoff bigram (6.1 million) and trigam (9.4 million) models and a compacted trigram model which has 16% lower perplexity and which is 60% smaller than the bigram model. We use the SMART IR system, developed for text-based retrieval using the vector space model. SMART tokenizes the transcribed audio, removes common stop words (e.g. *the*, *of*), normalizes word variants (e.g. *persons* → *person*), and weights term occurrences based upon document length and word frequency. SMART scores for 'hits' within a document are made available to the user interface, as well as the ranking of documents retrieved for a query. Two user interfaces currently exist for SCAN, a speech-based interface, and a text-based interface. Both of these interfaces reflect lessons learned from our earlier audio browsing experiments with simpler prototypes and a voicemail browsing and retrieval task.

2 Audio-Based Browsing and Retrieval

In recent years, various systems have been built to enable capture and browsing of spoken conversational data from meetings and recorded lectures (Hindus, Schmandt, and Horner, 1993; Kazman et al., 1996; Moran et al., 1997; Wolf, Rhyne, and Briggs, 1992; Whittaker, Hyland, and Wiley, 1994), and personally dictated information (Degen, Mander, and Salomon, 1992; Stifelman et al., 1993). Other systems allow search of multimedia archives of television programmes (Hauptmann and Witbrock, 1997; Shahraray, 1995) and videomail (Jones et al., 1996). While extensive evaluations of this technology remain to be carried out, naturalistic studies of audio browsing systems demonstrate their effectiveness in helping users produce accurate meeting summaries (Moran et al., 1997; Whittaker, Hyland, and Wiley, 1994; Wilcox, Schilit, and Sawhney, 1997). These and other studies also showed that indexed audio produces more accurate recall, although users may take longer to retrieve information (Kazman et al., 1996; Whittaker, Hyland, and Wiley, 1994). Several factors that may influence browsing behavior have been identified: (a) familiarity with subject matter: knowledgeable users are more likely to skip portions of the audio record when replaying (Moran et al., 1997) and they generate more effective queries when searching the record (Kazman et al., 1996); (b) type of retrieval task: audio search behaviors differ when users are trying to summarize as opposed to extract verbatim information from the audio record (Moran et al., 1997; Whittaker, Hyland, and Wiley, 1994); (c) presence and type of audio indices provided: cue utility is esoteric, with different users relying on different types of cue (Kazman et al., 1996); (d) availability of segmental information: users find it easier to navigate the record when structural information is provided (Arons, 1994). However, these studies also identify severe difficulties that users experience with speech browsing and search which may compromise the utility of these systems. The first problem is navigational: users often report losing track of the current audio context (Stifelman, 1996; Arons, 1994), and being unable to determine the sequence and structure of different elements of the audio record (Gould, 1983; Haas and Hayes, 1986). A second set of problems concern search: users seem to be poor at generating effective key word search queries, and find it hard to exploit system-generated key word indices. These problems are exacerbated when search material is unfamiliar (Kazman et al., 1996).

2.1 The Experiment Design

In our first set of experiments we focussed on identifying users' own search strategies when given a set of tasks involving access to a relatively small audio database and two relatively impoverished GUI interfaces to that database. More specifically we wanted to first identify the strategies users employ to browse and search audio — e.g., how do users find information in audio? Do they sample small segments or listen to large chunks? Second, we wanted to investigate the factors affecting these strategies — e.g., do users search familiar information differently from novel material and if so how? Is their search strategy different when they are looking for verbatim rather than summary information? Does providing segmental information aid search significantly? Do other kinds of cues or indices promote effective search? Third, we hoped to explore users' memory and mental models of audio and investigate the re-

lationship between memory and search strategies — do users with more accurate models search audio more effectively, and what promotes good memory?

We based our experiments on findings from a naturalistic study of over 800 voicemail users, in which we identified a set of strategies people used to access a real audio archive, and documented the problems users experience in accessing that archive (Hirschberg and Whittaker, 1997; Whittaker, Hirschberg, and Nakatani, 1998). In our laboratory experiments we focussed first on how access is affected by two factors, task type and familiarity of material. While previous research has suggested that these factors affect browsing, no detailed evaluation has been done. Second, we investigated the impact of two browser features, topic structure and play duration. Although these features have been implemented in previous browsers, their impact on browsing and their interaction with task and familiarity has not been systematically tested. Our hypotheses were that a) search efficiency (i.e. number of search operations and search time) depends on the amount of speech information users must access: summary tasks requiring access to an entire topic will less efficient than search for two specific facts, which in turn will be less efficient than search for one fact; b) familiar material will elicit more efficient search; c) providing information about where topics begin will increase the efficiency of search; and, d) short duration fixed play intervals will be used for identifying relevant topics, whereas longer fixed play durations will be used for search within a topic.

Fourteen people were given a speech archive, consisting of eight voicemail messages, or **topics**, appended together in one audio file 236.3 seconds long. We chose this domain for our study, both because we were interested in the specific application, and because Voicemail message retrieval is an example of a real application of audio search and retrieval, which we felt would be familiar to our users. Users accessed the archive to answer sixteen questions about the eight topics. These questions were based on retrieval tasks identified as common in our naturalistic study of voicemail users. There were three types of task: Four questions required users to access one specific fact, e.g. a date or phone number from a topic (**1fact**), a further four required access of two such facts (**2fact**), and eight questions required users to reproduce the gist of a topic (**summary**).

The first eight questions required users to access each of the eight topics once, and questions 9 through 16 required each topic to be accessed again. To investigate the effects of familiarity we compared users' performance on the first eight versus the second eight of the sixteen questions.

Users were given one of two GUI browsers: **basic** and **topic**. These are shown in Figure 2. Both

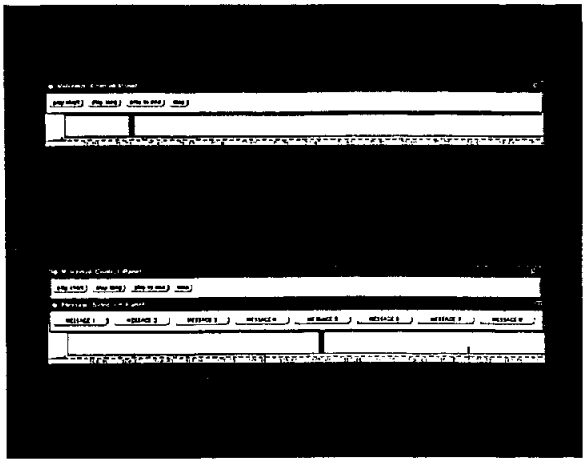

Figure 2: Basic and Topic Audio Browsers

browsers represent the entire speech archive as a horizontal bar and permit random access to it: users can select any point in the archive and play from that point (e.g. inserting the cursor halfway across the bar begins play halfway through the archive). For both browsers, users then select one of three play durations: *play short* (3 seconds), *play long* (10 seconds) and **play to end** (unrestricted play until play is manually halted by the user). The **topic browser** further allows the user to select a given topic by serial position (e.g. topic, or, message 1); play will then begin at the start of that topic/message.

We used a simple GUI for our initial experiments, rather than testing a complex set of possible search features, for three reasons: First, data on accessing a real speech archive indicate that even highly experienced users make little use of sophisticated features such as scanning, speed up/slow down, or jump forward/back (Hirschberg and Whittaker, 1997). Second, informal evaluations of complex speech UIs reveal that advanced browsing features are often not well understood by users, and do not necessarily improve search (Arons, 1994; Hauptmann and Witbrock, 1997). Given the unclear benefits of complex features, we wanted to establish baseline data for speech retrieval using a simple prototype. Finally, the features we tested will most likely be part of any browsing interface, and thus are of general interest.

Users were given 5-10 minutes on practice tasks before the experiment. After it, we gave users a memory test, asking them to recall the content, name of caller and serial position of each topic. We then administered a questionnaire eliciting reactions to browser features and comments about the tasks. We logged the number and type of each play operation, duration and location of played speech within the archive, and time to answer each question. The results for each hypothesis follow and all differences discussed are statistically significant at $p < 0.05$, using ANOVA.

2.2 Experimental Results

As we had expected, **1fact** tasks were answered more efficiently than both other tasks (see Table 1). However, contrary to expectations, **summary** was more efficient than **2fact**, despite requiring access to more information. The results indicate that performance depends both on the type and the amount of information users must access. User comments revealed why **2fact** were so difficult: with summaries it was possible to remember several pieces of approximate information. **2fact** questions required complex navigation within topic and the additional precision required to retain verbatim information often meant that users forgot one fact while searching for the second. They then found it hard to relocate the fact they had just forgotten. The user logs reveal problems of forgetting and relocating prior facts. In the course of answering each **2fact** usrquestion users actually played the two target facts a combined total of 7.9 times. In contrast target facts for **1fact** tasks were only accessed 1.5 times and topics 2.9 times for **summary** tasks.

As we had suspected, in general, familiar material elicited more efficient search. To investigate more deeply just **how** this effect was produced, we then separated overall search operations into: the identification of the relevant topic and the actual extraction of the information required to complete the task, i.e., finding the answer within the target topic. We then found that familiarity only improved the speed of topic identification, but had no effect on information extraction once the relevant source had been identified.

Users made frequent use of topic boundary information. Although random access was available with the topic browser, users only employed it for 33% of their access operations. Furthermore, users' comments about the topic boundary feature were highly positive. Despite this positive feedback however, we found that topic-based access seemed less efficient than random access: users with access to topic delimiters took more operations although less time to answer questions than other users. Why might this counter-intuitive result have occurred? Post-hoc tests showed that topic browser users had worse memory for the eight topics than simple browser users. Users of the basic browser reported making strenuous efforts to learn a mental model of the archive. In contrast, reliance on topic structure may permit topic browser users never to do so.

Play duration behavior was independent of whether search was within or outside topic. Furthermore, there was little use of either of the fixed play operations: all users preferred unrestricted play. In the final questionnaire, users reported that fixed duration options reduced their comprehension by truncating topic playback in unpredictable places. They preferred the greater control of unrestricted play, even though this meant the overhead of stopping play explicitly.

From these experiments we conclude, first, that users were much better at comprehending the overall structure of the archive, including the order and gist of topics, than they were at navigating more locally, within a given topic, to find particular pieces of information. They were unable, for example, to relocate previously accessed information within topic for **2fact** tasks, and showed no familiarity effects for search within topic. Second, our sampling results suggest that users overwhelmingly reject fixed duration *skims* of salient speech information, when given an alternative more within their control. Instead of fixed interval skimming, users prefer to access salient speech by controlling the precise playback duration themselves, even though this may involve more effort on their part to start and stop play. And third, providing topic boundaries may be of limited value: although users all like this feature (and those who participated in the basic browsing condition specifically requested it), heavy use of such signposts may make it more difficult for users to learn the contents of the archive. It appeared that the segmentation provided was at too coarse a level of granularity to provide much additional navigational power; the general topic structure of the archive as a whole could be learned easily without it.

3 Segmentation and Transcription Aids to Audio Navigation

The results of our basic and topic browser studies led us to propose two further browser prototypes, providing different types of additional, signposting information that might be helpful in local, as well as global navigation tasks. Since our goal was to permit users to browse much larger databases than eight voicemail messages, we also suspect that increasing the size of the database might increase the importance of some form of topic segmentation.

The first browser we developed provided a more sophisticated notion of topic segment than simple message boundaries, and is shown in Figure 3.

In our early study of heavy voicemail users we had learned anecdotally that callers who are used to doing business via voicemail believe that they and other such callers typically leave their messages in certain standard ways, with return telephone numbers and names at the beginning and end of messages, for example, and with content arranged in somewhat predictable fashion. So we prepared hand labelings of our test voicemail messages, identifying the following parts within each message: greeting, "Hi, Jim"; caller identification, "It's Valerie from the Customer Care committee"; topic, "I'm calling about the meeting next week"; deliverables, "Can

Task	Number of Operations	Solution Time
1fact	2.4	23.0
2fact	4.1	37.6
summary	2.9 (F = 7.43)	32.3 (F = 11.7)
familiar	2.1	22.5
unfamiliar	4.1 (F = 35.5)	40.1 (F = 36.6)
topic	3.7	30.0
no topic	2.5 (F = 5.09)	32.5 (F = 6.60)

Table 1: Effects of Task, Familiarity and Topic Structure on Retrieval Efficiency, with Relevant F ANOVA Values

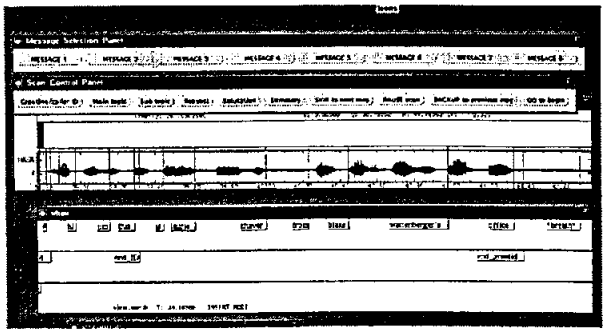

Figure 3: Voicemail Structural Browser

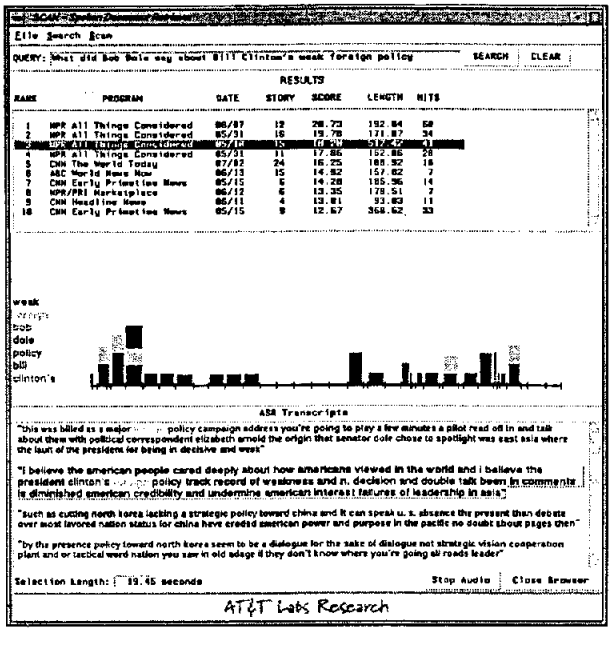

Figure 4: Text-Aided Audio Browser

you call Joan and make sure she'll have the numbers by then?"; and closing "Bye now." While we have tested this interface only informally, the addition of semantic categories as signposts to browsing through a series of messages seems much more useful than simply iterating through messages by start of message. A browse through caller identifying phrases, for example, quickly identifies messages by caller, while browsing through topics or deliverables serves the same function by topic. And playing greeting, caller id, topic, deliverables, and closing provides a very effective summary of many message. Of course, even this outwardly simple identification of topic structure is beyond the capability of existing technology. However, we are currently collecting and annotating a voicemail corpus with the goal of adding this type of structural browsing capability to the retrieval capabilities provided by our ASR/IR search engine.

The second browser we developed in order to experiment with new types of navigational aids to audio browsing makes use of the (errorful) ASR transcription produced in order to carry out IR in our system. This browser is depicted in Figure 4. The text-aided browser is implemented for the SDR Broadcast News corpus and consists of three main components: a programs overview window, a speech feedback window, and a player window. The programs overview presents the results of the IR search on the corpus in the form of a list of stories ranked in order of their relevance to a text input query. The top ten most relevant stories are displayed, in this version of the interface. For each story, the title of the program from which the story comes, the date of the broadcast, and all instances of keywords in the story that were deemed relevant to the query are displayed. Clicking on one of the program/story buttons loads the corresponding speech into the speech feedback window, along with a time-aligned cursor, which shows the location of the story in the speech stream. The player window then provides controls for navigation and play within the displayed program, permitting the following functionality: a play button with plays from the point selected in the speech feedback window; a stop play button; a move-to-beginning button; buttons which skip forward in the speech by intonational phrase or larger intonationally defined units and buttons which skip back-

ward in the same units. We have devised a series of tasks appropriate to the broadcast news domain but similar to the tasks used in our voicemail study, and will use this interface to test the utility of automatically derived transcription and keyword identification, as well as acoustically identified prosodic units, in aiding local navigation.

4 Discussion

A central problem with current access to large audio databases is the need to listen to large amounts of relevant data; the human eye skims much more quickly than is possible for the human ear to do. Also, when skimming text, humans typically are provided with many conventional orthographic and formatting guides, such as headings and paragraphs. Our study of audio browsing in even a small audio corpus demonstrates that, while some kind of navigational aids seem necessary to provide the context which permits successful navigation, obvious signposts such as topic/message boundaries may be less helpful than users expect them to be and perhaps even counter-productive to users acquiring a basic understanding of their data. Given this result, we are exploring alternatives to simple topic markers, including semantic structural information, potentially errorful transcription and key word retrieval, and acoustic segmentation, particularly as a means of enhancing users' ability to extract the information they seek from the audio data that has been presented to them.

References

Arons, B. 1994. *Interactively Skimming Speech.* Ph.D. thesis, MIT Media Lab.

Degen, L., R. Mander, and G Salomon. 1992. Working with audio: Integrating personal tape recorders and desk-top computers. In *Human Factors in Computing Systems: CHI '92 Conference Proceedings*, pages 413–418.

Gould, J. 1983. Human factors challenges: The speech filing system approach. *ACM Transactions on Office Information Systems*, 1(4), October.

Haas, C. and J. Hayes. 1986. What did i just say? reading problems in writing with the machine. *Research in the Teaching of English*, 20(1).

Hauptmann, A. and M. Witbrock. 1997. News-on-demand multimedia information acquisition and retrieval. In M. Maybury, editor, *Intelligent Multimedia Information Retrieval.* AAAI Press.

Hindus, D., C. Schmandt, and C. Horner. 1993. Capturing, structuring, and representing ubiquitous audio. *ACM Transactions on Information Systems*, 11:376–400.

Hirschberg, Julia and Steve Whittaker. 1997. Studying search and archiving in a real audio database. In *Proceedings of the AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video and Audio Corpora*, Stanford, March. AAAI.

Jones, G. J. F., J. T Foote, K. Sparck Jones, and S. J. Young. 1996. Retrieving spoken documents by combining multiple index sources. In *Proceedings of SIGIR 96*, Zurich, August. ACM.

Kazman, R., R. Al Halimi, W. Hunt, and M. Mantei. 1996. Four paradigms for indexing video conferences. *IEEE Multimedia*, 3(1):63–73.

Moran, T. P., L. Palen, S. Harrison, P. Chiu, D. Kimber, S. Minneman, W. Van Melle, and P. Zellweger. 1997. "i'll get that off the audio": A case study of salvaging multimedia meeting records. In *Human Factors in Computing Systems: CHI '97 Conference Proceedings*, pages 202–209.

Shahraray, Behzad. 1995. Scene change detection and content-based sampling of video sequences. In Robert J. Safranek and Arturo A. Rodriguez, editors, *Proceedings of the SPIE Conference on Digital Video Compression: Algorithms and Technologies*, February.

Stifelman, L. 1996. Augmenting real-world objects: A paper-based audio notebook. *Human Factors in Computing Systems: CHI '96 Conference Companion*, pages 199–200.

Stifelman, L., B. Arons, C. Schmandt, and E. Hulteen. 1993. Voicenotes: A speech interface for a hand-held voice notetaker. In *Human Factors in Computing Systems: CHI '93 Conference Proceedings*, pages 179–186.

Whittaker, S., P. Hyland, and M. Wiley. 1994. Filochat: Handwritten notes provide access to recorded conversations. In *Human Factors in Computing Systems: CHI '94 Conference Proceedings*, pages 271–277, New York. ACM Press.

Whittaker, Steve, Julia Hirschberg, and Christine Nakatani. 1998. All talk and all action: strategies for managing voicemail messages. In *Human Factors in Computing Systems: CHI '98 Conference Proceedings*, Los Angeles.

Wilcox, L. D., B. N. Schilit, and N. Sawhney. 1997. Dynomite: A dynamically organized ink and audio notebook. In *Human Factors in Computing Systems: CHI '97 Conference Proceedings.*

Wolf, C., J. Rhyne, and L. Briggs. 1992. Communication and information retrieval with a pen-based meeting support tool. In *Proceedings of CSCW-92*, pages 322–329.