NOW YOU HEAR IT, NOW YOU DON'T: EMPIRICAL STUDIES OF AUDIO BROWSING BEHAVIOR

Christine H. Nakatani, Steve Whittaker, and Julia Hirschberg
AT&T Laboratories-Research
180 Park Avenue, Florham Park NJ 07932-0971, USA
email: {*chn,stevew,julia*}*@research.att.com*

ABSTRACT

We present several studies that investigate how people use audio documents and uncover new principles for designing audio navigation technology. In particular, we report on an ethnographic study of voicemail users, experimental studies of human voicemail processing and the design of a new structural browser that embodies principles learned from the forementioned empirical studies. Specifically, our studies show that the reinforcement of audio memory and appropriate definition of content-based playback units are important properties of interfaces suited to human audio processing behaviors.

1. INTRODUCTION

This paper presents several studies investigating how people use audio documents, such as voicemail and recordings of lectures and meetings. Our empirical and experimental findings suggest new directions for the design of audio navigation technology. Below, we first report on an ethnographic study exploring the behaviors and needs of users of current voicemail technology. To constrain design choices for better technology, we then study how people navigate through audio and how they perform basic information processing tasks on a voicemail corpus. Observations and analyses from the user experiments lead to new principles of design for audio document interfaces. These principles are embodied in a prototype structural audio browser that we propose as a new interface for voicemail archives.

The research in this paper is part of a larger project at AT&T Labs called SCAN (Speech and Content-based Audio Navigation). The SCAN project addresses issues of pre-processing speech for presentation in interfaces, methods for retrieving speech documents, and the design of speech document applications and the interfaces themselves, which is the focus of this paper. The voicemail domain was chosen as one area of investigation within SCAN because voicemail is a ubiquitous and heavily used speech technology, yet the usability of voicemail and the relatively slow development of voicemail technology remain frustrating to those who rely on it as an important means of spoken communication.

2. PREVIOUS WORK

In recent years, various systems have been built to enable capture and browsing of spoken conversational data from meetings and recorded lectures [6, 9, 10, 17, 15], and personally dictated information [2, 13]. Other systems allow search of multimedia archives of television programmes [5, 11] and videomail [8]. 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 [10, 15, 16]. These and other studies also showed that indexed audio produces more accurate recall, although users may take longer to retrieve information [9, 15]. 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 [10] and they generate more effective queries when searching the record [9]; (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 [10, 15]; (c) presence and type of audio indices provided: cue utility is esoteric, with different users relying on different types of cue [9]; (d) availability of segmental information: users find it easier to navigate the record when structural information is provided [1]. 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 [12, 1], and being unable to determine the sequence and structure of different elements of the audio record [3, 4]. 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 [9].

3. HOW PEOPLE USE VOICEMAIL

We carried out a naturalistic or ethnographic study of 782 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 [7, 14]. The study consisted of collecting and analyzing (1) server data and usage statistics for 21 days; (2) questionnaire data from 133 high volume users (i.e. people who received more than 10 messages per day) probing their strategies for retrieving, archiving and managing voicemail data, and the extent of their use of existing features and capabilities available in their voicemail system; and (3) interview data with 15 high volume users exploring questions of their technology use in depth.

The ethnographic study revealed that users encountered two major search problems: scanning and information extraction. The first kind of search, scanning, is necessary to relocate messages already received, or to quickly

overview the contents of the voicemail archive under time constraints, for example, in between meetings or in transit. The interview data indicated that voicemail messages contain complex information, and are on average 30 seconds to 2 minutes in length for all users. Some users attempt to memorize the serial position of messages. However, these users are in the minority as 76% of survey respondents report that "listening to each message in sequence" is their standard procedure for finding archived messages. When scanning, users rarely make use of advanced system functionality, such as access to header information or faster playback, but rather rely on strategies such as listening for a certain speaker's voice in the first few seconds of a message.

The second search problem involves extracting information from the relevant message once it has been identified. Users report this is a laborious process, involving repeated playing of message parts while information is transcribed or committed to memory. In fact, 72% of survey respondents report "almost always" taking written notes while listening to voicemail messages. Sometimes, listeners seek specific facts; at other times, they seek to be reminded of the gist of the message, especially when processing a voicemail archive after a short time period away. Again, when extracting information, users rarely make use of advanced system functionality, such as skipping ahead or backward (by 3 seconds), or slowing down the playback rate (e.g. when trying to write down notes). The essentially linear access methods to voicemail archives seem to lead to strategies of serial access during scanning, as well as serial, repeated listenings to specific voicemail messages during information extraction.

4. HUMAN PROCESSING OF AUDIO

The fact that users do not exploit advanced functionality in existing technology is a curious one, especially given that they deem voicemail to be a laborious if necessary information technology. To determine whether audio navigation functionality could be better designed to suit users' reported needs, we undertook an experimental study of voicemail processing by experienced voicemail users.

4..1 The Experimental Design

Fourteen users were given a set of tasks involving access to a relatively small audio database and two relatively simple, underspecified GUI interfaces to that database. Crucially, the interfaces allowed for random access and random playback of speech messages, as well as providing limited basic access and playback functions. One reason for keeping the interfaces as simple as possible was to allow users to evolve their own strategies for processing voicemail given unrestricted access and playback capabilities. Further, the ethnography indicated that even highly experienced users make little use of sophisticated features such as scanning, speed up/slow down, or skip forward/back [7]; independent informal evaluations of complex speech UIs reveal that advanced browsing features are often not well understood by users, and do not necessarily improve search [1, 5]. 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.

In the experimental design, 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. Similarly, the 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. 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**. Both browsers represent the entire speech archive as a rectangular strip 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 strip 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.

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.

4..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 and had 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** question 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.

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

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.

5. DESIGN OF NEW TECHNOLOGY

The user experiments provided important insights into human processing of audio. First, it seems that users familiarize themselves with audio archives and their contents by repeated listenings. Yet, in current interfaces they have little control over how they listen, confined to a tape player model of (mostly) linear playback. A good scanning interface would improve the user's memory of the audio by selective repetition of memorable parts of the audio. Second, counter to expectations, fixed duration play commands, such as playing a short segment or long segment, do not satisfy user's needs to sample audio or absorb audio contents respectively. Rather, especially when extracting information, users seem to prefer hearing coherent stretches of audio messages; if they are longer than need be, listeners simply tune out and tune in to the audio stream according to their task needs. Users reported that it would be highly desirable to have meaningful or important parts of the messages identified; in contrast, it did not seem meaningful to determine ahead of time the length of audio to be played, partly because message contents and structure could not be reliably committed to memory.

In sum, signposting information needs to be identified and reinforced, based on message content and not strictly temporal units. To explore different avenues for signposting, we developed a structural voicemail browser that uti-

lizes a more sophisticated, fine-grained notion of topic segment than simple message boundaries. The structural browser is shown in Figure 1.

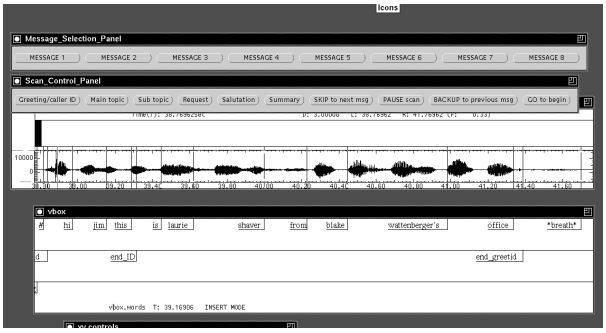

Figure 1: Voicemail Structural Browser

In our ethnography, we learned that 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 (ID)**, "It's Valerie from the Customer Care committee"; **topic**, "I'm calling about the meeting next week"; **deliverables**, "Can 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 ID phrases, for example, quickly identifies messages by caller, while browsing through topics or deliverables serves the same function by topic. And playing caller ID, topic and deliverables provides a very effective summary for most messages.

In addition, instead of providing fixed duration playback capabilities, we introduced playback *loops*, which iterate through the messages from the current location in the archive to the end, playing specific parts of each message. Playback loops for each type of message part and two combinations of message parts are provided, including *openings* (greetings, caller ID), *topics*, *deliverables*, *closings* and *summaries* (caller ID, topics, deliverables). The playback loops capture the kind of repeated serial access behaviors reported by users in the ethnography and observed in the experiments. Besides automating a new kind of listening strategy, they allow for efficient reinforcement of audio memory, organized along semantic dimensions.

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 automating this kind of structural browsing. Ongoing work is focused on applying segmentation algorithms to determine topic structure within messages, using transcriptions from an automatic speech recognizer as well as the acoustic-prosodic features of the speech directly. We are also exploring the generality of our findings by conducting related audio information processing experiments on an audio database of broadcast news.

References

[1] B. Arons. *Interactively Skimming Speech*. PhD thesis, MIT Media Lab, 1994.

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

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

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

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

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

[7] J. Hirschberg and S. Whittaker. 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 1997. AAAI.

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

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

[10] T. P. Moran, L. Palen, S. Harrison, P. Chiu, D. Kimber, S. Minneman, W. Van Melle, and P. Zellweger. "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, 1997.

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

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

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

[14] S. Whittaker, J. Hirschberg, and C. Nakatani. All talk and all action: strategies for managing voicemail messages. In *Human Factors in Computing Systems: CHI '98 Conference Proceedings*, Los Angeles, 1998.

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

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

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