Audio Browsing and Search in the Voicemail Domain

Julia Hirschberg and **Michiel Bacchiani** and **Phil Isenhour**
and Aaron Rosenberg and Larry Stead and Steve Whittaker and Gary Zamchick
AT&T Labs – Research
Florham Park NJ, USA
julia@research.att.com

Abstract

Increasing amounts of public, corporate, and private audio data are available electronically. However, they are limited in their usefulness by the lack of tools to browse and search them. In this paper we describe SCANMail, a system that employs automatic speech recognition, information retrieval, information extraction, and human computer interaction technology to allow users to browse and search their voicemail messages by content. SCANMail also provides note-taking capabilities as well as browsing and querying features. A CallerId server suggests caller names from existing caller acoustic models and is trained via user feedback. An Email server sends the original message plus its transcription to a mailing address specified in the user's profile.

1 Introduction

With decreasing storage costs, increasing amounts of public, corporate, and private audio, such as news and entertainment broadcasts, recorded audio conferences and focus groups, voicemail, are becoming available for search. But methods for searching audio corpora are far inferior to methods for searching text. Below, we describe a system for browsing and searching in a widely used speech application, voicemail.

This system follows a general approach to searching speech by content developed earlier at Cambridge University (Brown et al., 1994) for voicemail and extended to the broadcast news domain in the NIST TREC Spoken Document Retrieval effort (Garofolo, Auzanne, and Voorhees, 2000). Our current work employs new acoustic modeling techniques for a multi-media mail domain; uses information extraction strategies for locating key pieces of information in messages; proposes caller identification for messages based upon acoustic data; and develops and extensively tests graphical user interfaces to enhance these tasks. Our work is based upon a larger study of voicemail users, including 15 interviews, server data from 783 active users and a survey of 133 high volume users (Whittaker, Hirschberg, and Nakatani, 1998a), and experiments designed to identify problems in audio navigation (Whittaker, Hirschberg, and Nakatani, 1998b). This paper discusses corpus on which our system was trained, as well as the component parts of SCANMail.

2 The Corpus

The SCANMail training corpus was collected from 140 AT&T employees who volunteered their voicemail inboxes for the collection. Collection took place during a twelve-week period in early 1998. 105 hours of messages were collected, transcribed, and identified wherever possible as to caller, gender, age (adult/child), native/non-native speaker, and recording condition (e.g. cell phone). Some kinds of information was also labeled, to serve as training material for information extraction experiments, including greetings (e.g. "Hi how are you."), caller identification segments (e.g. "It's Molly."), tele-

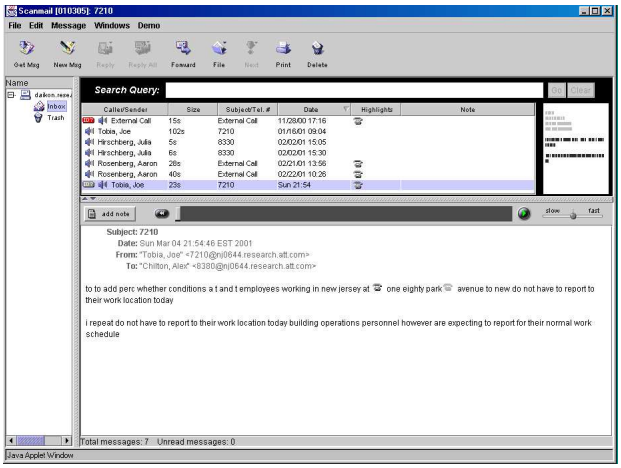

Figure 1: The SCANMail User Interface

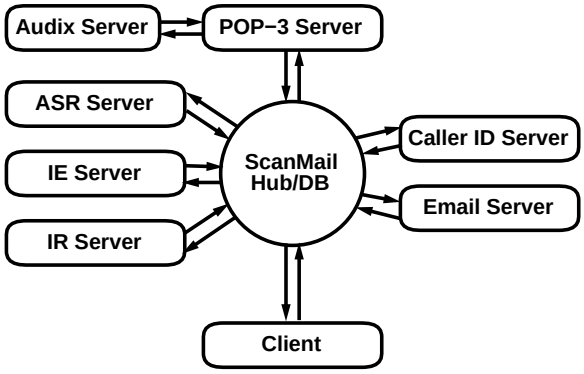

Figure 2: The SCANMail Architecture

phone numbers, times, dates, and closings (e.g. "Bye bye."). The final corpus, with duplicates (broadcast and forwarded messages) excluded, includes approximately 100 hours of speech, with 10,000 messages from approximately 2500 speakers. Most (about 90%) of the messages were recorded from regular handsets, and the rest from cellular and speaker-phones. The corpus is roughly gender balanced, with about 12% of the messages coming from non-native speakers. Mean duration of messages was 36.4 seconds and median was 30.0 seconds.

3 The SCANMail System

SCANMail uses automatic speech recognition (ASR), information retrieval (IR), information extraction (IE), and human computer interaction technology to enhance user access to their voicemail messages through a graphical user interface (GUI). Access to messages and information about them is presented to the user via a Java applet running under Netscape. Figure 1 shows the SCANMail GUI. Voicemail messages are retrieved from the Avaya voicemail system, *Audix*, via a POP3 server. Messages are stored in the SCANMail message store and processed by a number of SCANMail components. Figure 2 shows the architecture of the system.

New messages are processed initially by an ASR server, which transcribes them (shown in Figure 1), so that messages can be read or played, in whole or in part. The transcript is then indexed by the IR server, so that messages can subsequently be searched by content. The Email server sends the original message plus its ASR transcription to an email address specified in the user's profile. A CallerId server suggests a caller identifications by comparing the new message to acoustic models stored in its inventory for callers previously identified by the user as having left messages; users provide feedback on CallerId hypotheses so that the server can refine its initial models and create new ones. The SCANMail GUI provides access to all this information, as well as to the messages themselves, and to header information available from Audix itself or the PBX. The GUI also supports electronic note-taking capabilities as well as a variety of random access playing and querying features.

3.1 Automatic Speech Recognition

In SCANMail, messages are first retrieved from a voicemail server, then processed by the ASR server that provides a transcription for use by the IE, IR, Email and CallerId servers. The acoustic and language models of the recognizer, as well as the extraction and information retrieval techniques of the IE and IR servers are trained on 60 hours of the corpus described in Section 2. The ASR system itself uses a rescoring framework, where word graphs constructed by the baseline system are used as grammars for subsequent search passes. The system uses a 14,000 word vocabulary, automatically generated by the AT&T

(Beutnagel et al., 1999) Labs NextGen Text To Speech system. The language model, a Katz-style backoff trigram model, is trained on 700,000 words from the training set transcriptions. The ASR word-error rate in the current implementation, tested on a 40 hour test set, is 34.9%. The ASR server, running on a 667 MHz 21264 Alpha processor, currently produces the final transcripts in approximately 20 times real-time. Details of this system are presented in (Bacchiani, 2001).

3.2 Information Retrieval

Messages transcripts are next indexed by the IR server using the SMART IR (Salton, 1971; Buckley, 1985) engine, which is based on the vector space model of information retrieval. SMART generates weighted term vectors for the message transcriptions, after tokenizing them, removing words on its stop-list of common terms, and stemming the resulting text. When the IR server executes a query, query terms are similarly converted into weighted term vectors and vector inner-product similarity comparison is used to rank messages in terms of their relevance to the query. In the SCANMail GUI, a search window then presents these results, with query terms color coded in the query itself, in the transcript and in the message thumbnail representation, described below. Relevant messages are presented with the most likely matches appearing first in the header list of the search window. Figure 3 shows the result of the query "Contractor estimate" in the SCANMail client.

3.3 Information Extraction

Key information is identified in the ASR transcription by the IE server, which currently extracts likely phone numbers identified from the message body. At present, this is accomplished in our production system by recognizing digit strings and scoring them based on the sequence length. An improved extraction algorithm, trained on our hand-labeled voicemail corpus, employs a digit string recognizer combined with a trigram language model, to recognize strings in their lexical contexts, e.g. <word> <digit-string> <word>, and

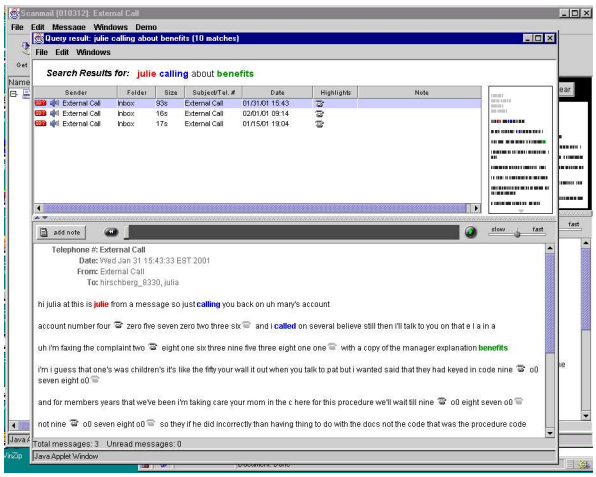

Figure 3: A SCANMail Query

has just been implemented in the development version of SCANMail. In both cases, the extracted information is made available to the user in several ways: First, a phone icon appears in the header of messages for which potential phone numbers have been extracted; a rollover feature allows users to view and play hypothesized numbers with their associated speech from the header window. Second, phone icons bracket possible phone numbers in the transcription. Future information to be extracted by the IE server includes names, dates, and times.

4 Caller Identification

The CallerID server suggests likely caller names in the header window by matching incoming messages against acoustic models trained from user-labeled calls. This capability is based on text independent speaker recognition techniques applied to voicemail. A user may label any message he/she has listened to at least a portion of, providing a caller name which is then used to create a new model for that caller or to augment an existing model. When the cumulative duration of labeled messages is sufficient (models with the same user-supplied label are shared across users), an initial caller model is constructed. Later messages received by the user will be processed and scored against this caller model as well as models for other callers the user

may have labeled messages for. If the best matching model's score for an incoming message exceeds a decision threshold, the CallerID server proposes a caller name to the GUI client. If there is no PBX-supplied identification (i.e. no caller name supplied from the owner of the extension for calls internal to the PBX), the CallerId hypothesis is presented in the message header, for acceptance or correction by the user. if there **is** a PBX-supplied identification, the CallerId hypothesis appears as the first item in the user's *contact menu*, together with all previously identified callers for that user.

The callers selected by the user for identification are referred to as *ingroup*; all other callers are *outgroup*. Three types of CallerID errors are possible: 1) An outgroup caller can be identified as ingroup: *outgroup acceptance*. 2) An ingroup caller can be identified as another ingroup caller: *ingroup confusion*. 3) An ingroup caller can be labelled as "unknown": *ingroup rejection*. A subset of the training corpus was used to evaluate CallerID performance. With decision thresholds set to maintain outgroup acceptance at the relatively low level of 2.7%, ingroup rejection is 11.5% and ingroup confusion is 1.2% for a 20-caller ingroup. Details of the CallerId process and performance evaluation are described in (Rosenberg et al., 2000).

5 The User Interface

The ScanMail GUI provides users with access to messages and information about them. The GUI shows message headers including: callerid, time and date, length in seconds, and (if available) telephone icons indicating extracted telephone numbers, as well as the first line of any attached note. Users also are given a thumbnail image of the current message and its ASR transcription. Any note attached to the current message is also displayed when the message is selected. A search panel allows users to search the contents of their mailboxes by typing in any text query (see Figure 3), with results presented in a new search window, with keywords color-coded in the query, transcript, and thumbnail. The GUI also supports various audio playing operations, including playing the entire message or "audio paragraphs" (*paratones*) selected from the transcript. Users can also highlight regions of the transcript and play the segment of the audio message corresponding to the selected text. Audio play speed can be modified by the user, who can speed up or slow down messages using a slider bar.

6 System Evaluation

To decide whether SCANMail is superior for voicemail access to current touchtone phone interfaces, we conducted a user study comparing SCANMail to over-the-phone Audix access. Eight subjects performed a set of fact-finding, message identification, and summarization tasks on artificial mailboxes of twenty messages each, using either SCANMail or Audix access. Each subject used both systems, with the order of system type, task, and inbox systematically varied. For the fact-finding task, users were asked to find two facts which appeared in some message in the inbox, such as the room number of a meeting and the title of a talk they had been asked to give. For the message identification task, they were asked to identify the most relevant message to answering a particular question, such as how to replace a lost badge, when there were multiple messages relevant to this question. For the summarization task, they were asked to summarize a particular message, e.g. to summarize directions to an off-site meeting. All eight subjects had used the regular voicemail system, although none had previously seen SCANMail. To address this discrepancy somewhat, they were given brief tutorials in both the voicemail system and in SCANMail at the beginning of the experiment.

We initially hypothesized that SCANMail would permit users to perform all tasks faster and more correctly than the regular voicemail system. We expected greater advantages for the fact-finding and message identification tasks, since these required users to find particular messages, as well as to locate information in them. Thus, SCANMail's search capabilities should provide an improvement over

standard voicemail serial search.

We collected both objective and subjective measures. The objective measures included time to completion of task, quality of answer (hand-scored by the experimenters), and a combined measure of "quality of answer/time". The subjective measures were collected from questionnaires subjects filled out after completion of each task and at the end of the experiment. Questions were asked about how time-consuming the task was felt to be, how easy, and how useful the interface was; subjects were also asked to rate each feature of the interface with respect to the preceding task and over all.

Results of our experiments confirmed advantages for SCANMail for both fact-finding and message identification tasks in the combined quality/time measure ($p < .05$). Solutions for the fact-finding task were also faster with SCANMail ($p. < .01$). And there was a trend toward a higher combined score across all task types with SCANMail ($p < .09$). On subjective measures, subjects rated SCANMail higher than regular voicemail access on all measures. Normalized performance scores were higher when subjects employ IR searches that were successful (i.e. the queries they choose contained words correctly recognized by the recognizer) ($p < .05$). Normalized performance scores were also higher for subjects who listened to less audio ($p < .05$) – presumably because they relied more upon the ASR transcripts. SCANMail's search capability, its transcripts, and the playbar were its most highly rated features; while the note facility and the thumbnail representation were not found to be useful for these tasks. In observing subjects' performing the tasks, we informally noted that SCANMail's search capability could be misleading: When subjects relied upon its accuracy too heavily, they sometimes assumed that they had found all relevant documents after a SCANMail search, when in fact some were **not** retrieved (since query terms in the relevant messages had not been recognized correctly by the ASR server), leading to a failure to find desired information. Similarly, when subjects trusted their reading of the (inaccurate) ASR transcripts, they sometimes missed crucial but unrecognized information.

So, it appears that indeed SCANMail does indeed offer some increase in efficiency and a significant increase in user perceived utility over regular voicemail access. A trial of eighteen "friendly" users accessing their own voicemail via the prototype has recently been completed, using a version of the client with modifications to access functionality which were suggested by our subject users. A total of 37 users currently employ SCANMail to access their voicemail.

7 Discussion

SCANMail integrates speech, computational linguistics, information retrieval, and human-computer interaction technologies and research efforts to provide new capabilities for browsing and searching audio corpora. Our current prototype system is now used by "friendly" users at AT&T Labs to access their voicemail by content via a GUI interface. In this system, messages are processed by ASR, IR, IE, and CallerId servers to produce transcriptions, searchable indices, extracted phone numbers, and hypothesized caller identification. The GUI allows a variety of random access play and search capabilities. We currently have alternative access interfaces, including an over-the-phone interface and PDA access, in prototype.

References

Bacchiani, Michiel. 2001. Automatic transcription of voicemail at AT&T. In *Proceedings of ICASSP–01.*

Beutnagel, M., A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. 1999. The AT&T Next-Gen TTS system. In *Proceedings of the Join Meeting of ASA,EAA, and DEGA*, Berlin, March. Paper No. 2aSCa4, J. Acoust. Soc. Amer. 105 (2) 1030 (A).

Brown, M. G., J. T. Foote, G. J. F. Jones, K. Sparck Jones, and S. J. Young. 1994. Video mail retrieval by voice: An overview of the cambridge/olivetti retrieval system. In *Proceedings of the 2nd ACM International Workshop*

on Multimedia Data Base Management, pages 47—55, San Francisco, October.

Buckley, Chris. 1985. Implementation of the SMART information retrieval system. Technical Report TR85-686, Department of Computer Science, Cornell University, Ithaca, NY 14853, May.

Garofolo, J. S., C. G. P. Auzanne, and E. M. Voorhees. 2000. The TREC Spoken Document Retrieval track: A success story. In *Proceedings of RIAO 2000: Content-Based Multimedia Information Access*, volume 1, pages 1–20, Paris.

Rosenberg, A., S. Parthasarathy, J. Hirschberg, and S. Whittaker. 2000. Foldering voicemail messages by caller using text independent speaker recognition. In *Proceedings of the Sixth International Conference on Spoken Language Processing*, Beijing.

Salton, Gerard, editor. 1971. *The SMART Retrieval System—Experiments in Automatic Document Retrieval*. Prentice Hall Inc., Englewood Cliffs, NJ.

Whittaker, Steve, Julia Hirschberg, and Christine Nakatani. 1998a. All talk and all action: strategies for managing voicemail messages. In *Proceedings of CHI '98*, Los Angeles.

Whittaker, Steve, Julia Hirschberg, and Christine Nakatani. 1998b. Play it again: a study of the factors underlying speech browsing behavior. In *Proceedings of CHI '98*, Los Angeles.