SCANMail: Browsing and Searching Speech Data by Content

Julia Hirschberg[1], *Michiel Bacchiani*[1], *Don Hindle*[2], *Phil Isenhour*[4],
Aaron Rosenberg[1], *Litza Stark*[3], *Larry Stead*[1], *Steve Whittaker*[1], *and Gary Zamchick*[1]

AT&T Labs – Research[1], AnswerLogic[2], University of Delaware[3], Virginia Tech[4]

{julia,michiel}@research.att.com, dhindle@answerlogic.com, isenhour@vt.edu,
aer@research.att.com, litza@udel.edu, {lstead,stevew}@research.att.com, zamchick@att.com

Abstract

Increasing amounts of public, corporate, and private audio data are available for use, but limited in usefulness by the lack of tools to permit their browsing and search. In this paper, we describe SCANMail, a system that employs automatic speech recognition, information retrieval, information extraction, and human computer interaction technology to permit users to browse and search their voicemail messages by content through a graphical user interface interface. The SCANMail client also provides note-taking capabilities as well as browsing and querying features. A CallerId server also proposes caller names from existing caller acoustic models and is trained from 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 storage costs shrinking, increasing amounts of public, corporate, and private audio — news and entertainment broadcasts, recorded audio conferences and focus groups, voicemail — are available for search. But methods for searching audio corpora fall far short of text-based search techniques. Without similar tools for navigating speech data, people are unable to take advantage of spoken databases without laborious hand-indexing.

In this paper, we describe a system for browsing and searching in a widely used speech application, voicemail. We follow a general paradigm for audio search systems, developed earlier at Cambridge University [1] for voicemail and extended to the broadcast news domain in the NIST TREC Spoken Document Retrieval effort [2]. Our work extends these efforts by employing new acoustic modeling techniques for a multi-media mail domain; using information extraction strategies for locating key pieces of information in messages; proposing caller identification for messages based upon acoustic data; and developing and extensively testing interfaces to make this technology useful for potential consumers. 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 [3], and experiments designed to identify problems in audio navigation [4]. In this paper we describe the component parts of our SCANMail system and discuss results of experiments we have performed which compare it with standard over-the-phone voicemail access.

2. The SCANMail System

The SCANMail system employs automatic speech recognition (ASR), information retrieval (IR), information extraction (IE),

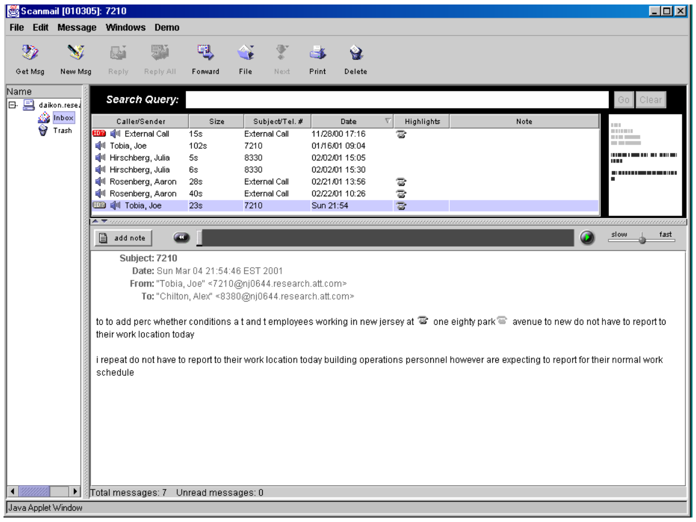

Figure 1: The SCANMail User Interface

and human computer interaction technology to allow users enhanced 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 a commercial voicemail system, *Audix*, an Avaya messaging system, via a POP3 server which polls the Audix voicemail server. Messages are then stored in the SCANMail message store and processed by a number of SCANMail components. Figure 2 shows the architecture of the system.

A new message is first processed by the ASR server, which produces a transcript of the message (shown in Figure 1), so that messages can be read or played, in whole or in part. The transcript is next 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. Additionally, a CallerId server proposes a caller identification by comparing the new message to acoustic models in its inventory which exist for callers previously identified as having left messages for this recipient; users are asked to provide feedback on CallerId hypotheses so that this 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 header information available from Audix itself or the PBX; it also supports electronic note-taking capabilities as well as a variety of random access playing and querying features.

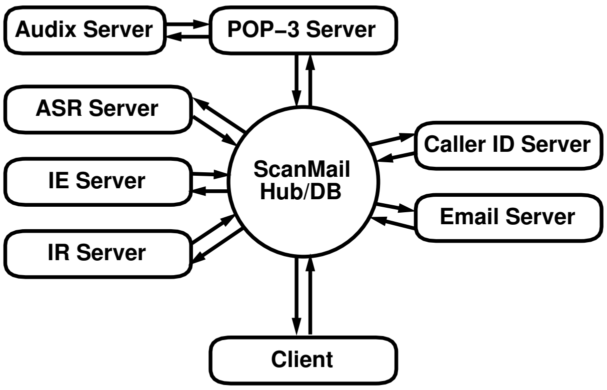

Figure 2: The SCANMail Architecture

3. The Training Corpus

The SCANMail training corpus was collected from voicemail messages received by 140 AT&T employees who volunteered their mailboxes for the collection. The collection period was a twelve-week period in early 1998. 105 hours 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). Certain types of information were also bracketed and labeled, to serve as training material for information extraction experiments, including greetings (e.g. "Julia hi."), caller identification segments (e.g. "It's Jane."), telephone numbers, times, dates, and closings (e.g. "Talk to you soon."). The final corpus, with duplicates (broadcast and forwarded messages) excluded, includes approximately 100 hours of speech, with 10,000 messages from approximately 2500 speakers. About 90% of the messages were recorded from regular handsets, the rest from cellular and speaker-phones. The corpus is approximately gender balanced. Approximately 12% of the messages were from non-native speakers. The mean duration of messages was 36.4 seconds; the median was 30.0 seconds.

4. Automatic Speech Recognition

In SCANMail, messages are first retrieved from a voicemail server, then processed by the ASR server that provides a transcription. The message audio and/or transcription are then passed to the IE, IR, Email, and CallerId servers. The acoustic and language model of the recognizer, and the IE and IR servers are trained on 60 hours of the corpus.

The ASR system uses a rescoring framework, where the word graphs constructed by the baseline system are used as grammars for subsequent search passes. This baseline system is a decision-tree based state-clustered triphone system with 8000 tied states. The emission probabilities of the states are modeled by 12 component Gaussian mixture distributions. The system uses a 14,000 word vocabulary, automatically generated by the AT&T [5] Labs NextGen Text To Speech system. The language model is a Katz-style backoff trigram trained on 700,000 words from the transcriptions of the 60 hour training set. The word-error rate of this system on a 40 hour test set is 34.9%.

Since the messages come from a highly variable source both in terms of speaker as well as channel characteristics, transcription accuracy is significantly improved by application of various normalization techniques, developed for Switchboard evaluations [6]. The ASR server uses gender-dependent models, Vocal Tract Length Normalization (VTLN) [7], Constrained Modelspace Adaptation (CMA) [8], Maximum Likelihood Linear Regression (MLLR) [9] and Semi-Tied Covariances (STC) [10] to obtain progressively more accurate acoustic models and uses these in a rescoring framework. In contrast to Switchboard, voicemail messages are generally too short too allow direct application of the normalization techniques. A novel message clustering algorithm based on MLLR likelihood [11] is used to guarantee sufficient data for normalization. The final transcripts, obtained after 6 recognition passes, have a word error rate of 28.7% – a 6.2% accuracy improvement. Gender dependency provides 1.6% of this gain. VTLN then additively improves accuracy with 1.0% when applied only on the test data and an additional 0.3% when subsequently applied with a VTLN trained model. The use of STC further improves accuracy with 1.2%. Finally CMA and MLLR provide additive gains of 1.5% and 0.6% respectively. A forced alignment of the audio against the final transcript provides word-level time marks for use in the GUI. A detailed analysis of the ASR performance on this task is provided in [12]. The ASR server, running on a 667 MHz 21264 Alpha processor, produces the final transcripts in approximately 20 times real-time.

5. Information Retrieval

Messages transcripts are indexed by the IR server using the SMART IR [13, 14] engine. SMART is based on the vector space model of information retrieval. It generates weighted term (word) vectors for the automatic transcriptions of the messages. SMART pre-processes the automatic transcriptions of each new message by tokenizing the text into words, removing common words that appear on its stop-list, and performing stemming on the remaining words to derive a set of terms, against which later user queries can be compared. When the IR server is used to execute a user query, the query terms are also converted into weighted term vectors. Vector inner-product similarity computation is then used to rank messages in decreasing order of their similarity to the user query. A new window presents search results, with query terms color coded in the query itself and in the transcript and thumbnail. Relevant messages are ranked from most to least likely to match the query. Figure 3 shows the result of the query "Contractor estimate" in the SCANMail client.

6. Information Extraction

Key information is extracted from the ASR transcription by the IE server, which currently extracts likely phone numbers identified in the message. At present, this is done 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>. Results are available to the user in several ways: 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. Phone icons also bracket hypothesized numbers in the ASR transcript. Future items to be extracted include names, dates, and times.

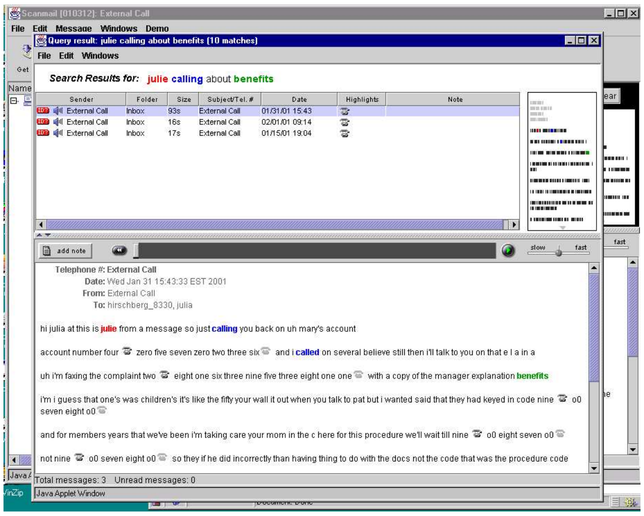

Figure 3: A SCANMail Query

7. Caller Identification

The CallerID server proposes caller names by matching messages against existing caller models; this module is trained from user feedback. The caller identification capability is based on text independent speaker recognition techniques applied to the processed speech in the voicemail messages. A user may elect to label a message he/she has reviewed with a caller name for the purpose of creating a speaker model for that caller. When the cumulative duration of such user-labeled messages is sufficient, a caller model is constructed. Subsequent messages will be processed and scored against this caller model and models for other callers the user may have designated. If the best matching model score for an incoming message exceeds a decision threshold, a caller name hypothesis is sent to the GUI client; if there is no PBX-supplied identification (i.e. 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 either accepting or editing by the user; if there is a PBX identification, the CallerId hypothesis appears as the first item in a user 'contact menu', together with all previously identified callers for that user. To optimize the use of the available speech data, and to speed model-building, caller models are shared among users. The callers selected by the user for identification are referred to as "ingroup". All other callers are "outgroup". There are three possible types of CallerID errors. An outgroup caller can be identified as ingroup: outgroup acceptance. One ingroup caller can be identified as another ingroup caller: ingroup confusion. 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 [15].

8. The User Interface

The ScanMail GUI provides 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 see a thumbnail image of the current message and its ASR transcription. Any note attached to the current message is also displayed. A search panel permits users to search the contents of their mailboxes by typing in any text query (see Figure 3). Results are 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. Finally audio playing speed can be customized, allowing messages to be speeded up or slowed down during playback.

9. Evaluation

To determine whether SCANMail is better for voicemail access than current touchtone phone interfaces, we conducted a user study comparing SCANMail to standard Audix access. Eight subjects performed a series of fact-finding, message identification, and summarization tasks on artificial mailboxes of twenty messages each, using either SCANMail or phone access. Each subject used both systems, with 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, but none had previously seen SCANMail. They were, however, given brief tutorials in both the voicemail system and in SCANMail at the beginning of the experiment.

We hypothesized that SCANMail would permit users to accomplish tasks faster and more correctly than the regular voicemail system. We expected there would be greater advantages for the fact-finding and message identification tasks, since these required users to locate messages, as well as to extract information from them. Thus, SCANMail's search capabilities should be an improvement compared with standard voicemail serial search. We collected both objective and subjective measures: 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". Subjective measures were gathered from a set of questionnaires subjects filled out after completion of each task and at the end of the experiment. They included questions 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.

There were advantages for SCANMail for both fact-finding and message identification tasks in the combined quality/time measure ($p < .05$). SCANMail also produced faster solutions for the fact-finding task ($p. < .01$). There was a trend toward a higher combined score across all task types ($p < .09$). On the 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 listen to less audio ($p < .05$) – presumably because they rely 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. We informally noted in observing subjects that SCANMail's search capability could be misleading: When subjects relied upon its accuracy, they sometimes assumed that they had found all relevant documents, when in fact some were **not** retrieved, leading to a failure to find desired information. Similarly, when subjects trusted the ASR transcript more than they should, they tended to miss crucial but unrecognized information.

We concluded that indeed SCANMail offers some increase in efficiency and a significant increase in perceived utility over regular voicemail access. A trial of ten "friendly" users accessing their own voicemail via the prototype is currently underway, with modifications to access functionality suggested by our subject users. A larger trial of the system is also being prepared, for more extensive testing of SCANMail use over time.

10. Discussion

The SCANMail system 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, in a 'friendly' trial, allows users access to their voicemail by content via a GUI interface. 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. Future research includes expanding the range of items to be extracted from transcripts, automatic message gisting, and new interfaces for over-the-phone and PDA access.

Acknowledgements

The authors would like to thank Mark Jones, Andrej Ljolje, S. Parthasarathy, Fernando Pereira, and Amit Singhal for their help in developing SCANMail and the underlying technology it employs.

11. References

[1] M. G. Brown, J. T. Foote, G. J. F. Jones, K. S. Jones, and S. J. Young, "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*, (San Francisco), pp. 47—55, October 1994.

[2] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC Spoken Document Retrieval track: A success story," in *Proceedings of RIAO 2000: Content-Based Multimedia Information Access*, vol. 1, (Paris), pp. 1–20, 2000.

[3] S. Whittaker, J. Hirschberg, and C. Nakatani, "All talk and all action: strategies for managing voicemail messages," in *Proceedings of CHI '98*, (Los Angeles), 1998.

[4] S. Whittaker, J. Hirschberg, and C. Nakatani, "Play it again: a study of the factors underlying speech browsing behavior," in *Proceedings of CHI '98*, (Los Angeles), 1998.

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

[6] *Proceedings of the Speech Transcription Workshop, University of Maryland*, May 2000.

[7] T. Kamm, G. Andreou, and J. Cohen, "Vocal tract normalization in speech recognition: Compensating for systematic speaker variability," in *Proceedings of the 15th Annual Speech Research Symposium*, (Johns Hopkins University, Baltimore, MD), pp. 161–167, 1995.

[8] M. J. F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, pp. 75–90, 1998.

[9] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, pp. 171–185, 1995.

[10] M. J. F. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 7, no. 3, 1999.

[11] M. Bacchiani, "Using maximum likelihood linear regression for segment clustering and speaker identification," in *Proceedings of the Sixth International Conference on Spoken Language Processing*, vol. 4, (Beijing), pp. 536–539, 2000.

[12] M. Bacchiani, "Automatic transcription of voicemail at AT&T," in *Proceedings of ICASSP–01*, 2001.

[13] G. Salton, ed., *The SMART Retrieval System—Experiments in Automatic Document Retrieval*, Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

[14] C. Buckley, "Implementation of the SMART information retrieval system," Tech. Rep. TR85-686, Department of Computer Science, Cornell University, Ithaca, NY 14853, May 1985.

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