SCANMail: Audio Navigation in the Voicemail Domain

Michiel Bacchiani  Julia Hirschberg  Aaron Rosenberg  Steve Whittaker
Donald Hindle       Phil Isenhour       Mark Jones      Litza Stark
Gary Zamchick

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

AT&T Labs – Research
180 Park Avenue
Florham Park, NJ 07932-0971, USA

ABSTRACT
This paper describes SCANMail, a system that allows users to browse and search their voicemail messages by content through a GUI. Content based navigation is realized by use of automatic speech recognition, information retrieval, information extraction and human computer interaction technology. In addition to the browsing and querying functionalities, acoustics-based caller ID technology is used to propose caller names from existing caller acoustic models trained from user feedback. The GUI browser also provides a note-taking capability. Comparing SCANMail to a regular voicemail interface in a user study, SCANMail performed better both in terms of objective (time to and quality of solutions) as well as subjective objectives.

1. INTRODUCTION
Increasing amounts of public, corporate, and private audio present a major challenge to speech, information retrieval, and human-computer interaction research: how can we help people to take advantage of these resources when current techniques for navigating them fail far short of text-based search methods? In this paper, we describe SCANMail, a system that employs automatic speech recognition (ASR), information retrieval (IR), information extraction (IE), and human computer interaction (HCI) technology to permit users to browse and search their voicemail messages by content through a GUI interface. 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 ASR transcription to a mailing address specified in the user’s profile. The SCANMail GUI also provides note-taking capabilities as well as browsing and querying features. 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.

2. SYSTEM DESCRIPTION
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 a 100 hour voicemail corpus, transcribed and hand labeled for telephone numbers, caller names, times, dates, greetings and closings. The corpus includes approximately 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 and approximately 12% of the messages were from non-native speakers. The mean duration of the messages was 36.4 seconds; the median was 30.0 seconds.

2.1 Automatic Speech Recognition
The baseline ASR system is a decision-tree based state-clustered triphone system with 8k tied states. The emission probabilities of the states are modeled by 12 component Gaussian mixture distributions. The system uses a 14k vocabulary, automatically generated by the AT&T Labs NextGen Text To Speech system. The language model is a Katz-style backoff trigram trained on 700k 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 [9]. The ASR server uses Vocal Tract Length Normalization (VTLN) [5], Constrained Modelspace Adaptation (CMA) [3], Maximum Likelihood Linear Regression (MLLR) [6] and Semi-Tied Covariances (STC) [4] to obtain progressively more accurate acoustic models and use these in a rescoring framework. In contrast to Switchboard, voicemail messages are generally too short to allow direct application of the normalization techniques. A novel message clustering algorithm based on MLLR likelihood [1] 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. The ASR
Figure 1: The SCANMail User Interface
server, running on a 667 MHz 21264 Alpha processor, produces
the final transcripts in approximately 20 times real-time.

2.2 Information Retrieval

Messages transcripts are indexed by the IR server using the SMART
IR [8, 2] engine. SMART is based on the vector space model
of information retrieval. It generates weighted term (word) vec-
tors for the automatic transcriptions of the messages. SMART pre-
processes the automatic transcriptions of each new message by to-
kenizing 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.

2.3 Information Extraction

Key information is extracted from the ASR transcription by the
IE server, which currently extracts any phone numbers identified
in the message. Currently, this is done by recognizing digit strings
and scoring them based on the sequence length. An improved ex-
traction 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>.

2.4 Caller Identification

The CallerID server proposes caller names by matching mes-
 sage 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 mes-
sage he/she has reviewed with a caller name for the purpose of
creating a speaker model for that caller. When the cumulative du-
ration 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 incom-
ing message exceeds a decision threshold, a caller name hypothesis
is sent to the GUI client; if there is no PBX-supplied identifica-
tion (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 id’d
callers for that user. To optimize the use of the available speech
data, and to speed model-building, caller models are shared among
users. Details and a performance evaluation of the CallerID process
are described in [7].

2.5 Graphical User Interface

In the SCANMail GUI, users see message headers (callerid, time
and date, length in seconds, first line of any attached note, and
presence of extracted phone numbers) as well as a thumbnail and
the ASR transcription of the current message. Any note attached
to the current message is also displayed. A search panel permits
users to search the contents of their mailbox by inputting any text
query. Results are presented in a new search window, with key-
words color-coded in the query, transcript, and thumbnail.

2.6 User Studies

User studies compared SCANMail with a standard over-the-phone
voicemail access. Eight subjects performed a series of fact-finding,
relevance ranking, and summarization tasks on artificial mailboxes
of twenty messages each, using either SCANMail or phone access.
SCANMail showed advantages for fact-finding and relevance rank-
ting tasks in quality of solution normalized by time to solution, for
fact-finding in time to solution and in overall user preference. Nor-
malized performance scores are higher when subjects employ IR
searches that are successful (i.e. the queries they choose contain
words correctly recognized by the recognizer) and for subjects who
listen to less audio and rely more upon the transcripts. However, we
also found that SCANMail’s search capability can be misleading,
causing subjects to assume that they have found all relevant doc-
uments when in fact some are NOT retrieved, and that when sub-
jects rely upon the accuracy of the ASR transcript, they can miss
 crucial but unrecognized information. A trial of 10 friendly users
is currently underway, with modifications to access functionality
suggested by our subject users. A larger trial of the system is be-
ing prepared, for more extensive testing of user behavior with their
own mailboxes over time.

Acknowledgements

The authors would like to thank Andrej Ljolje, S. Parthasarathy,
Fernando Pereira, and Amit Singhal for their help in developing
this application.

3. REFERENCES

[1] M. Bacchiani. Using maximum likelihood linear regression
for segment clustering and speaker identification. In
Proceedings of the Sixth International Conference on Spoken
Language Processing, volume 4, pages 536–539, Beijing,
2000.

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

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

[4] M. J. F. Gales. Semi-tied covariance matrices for hidden
markov models. IEEE Transactions on Acoustics, Speech,
and Signal Processing, 7(3), 1999.

[5] 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, pages 161–167, Johns
Hopkins University, Baltimore, MD, 1995.

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

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

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

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