SCANMail: a voicemail interface that makes speech browsable, readable and searchable

Steve Whittaker, Julia Hirschberg, Brian Amento, Litza Stark, Michiel Bacchiani, Philip Isenhour, Larry Stead, Gary Zamchick and Aaron Rosenberg
AT&T Labs-Research,
180 Park Avenue, Florham Park, NJ 07932, USA
stevew@research.att.com

ABSTRACT

Increasing amounts of public, corporate, and private speech data are now available on-line. These are limited in their usefulness, however, by the lack of tools to permit their browsing and search. The goal of our research is to provide tools to overcome the inherent difficulties of speech access, by supporting visual scanning, search, and information extraction. We describe a novel principle for the design of UIs to speech data: *What You See Is Almost What You Hear (WYSIAWYH).* In *WYSIAWYH*, automatic speech recognition (ASR) generates a transcript of the speech data. The transcript is then used as a *visual analogue* to that underlying data. A graphical user interface allows users to visually scan, read, annotate and search these transcripts. Users can also use the transcript to access and play specific regions of the underlying message. We first summarize previous studies of voicemail usage that motivated the *WYSIAWYH* principle, and describe a voicemail UI, SCANMail, that embodies *WYSIAWYH.* We report on a laboratory experiment and a two-month field trial evaluation. SCANMail outperformed a state of the art voicemail system on core voicemail tasks. This was attributable to SCANMail's support for visual scanning, search and information extraction. While the ASR transcripts contain errors, they nevertheless improve the efficiency of voicemail processing. Transcripts either provide enough information for users to extract key points or to navigate to important regions of the underlying speech, which they can then play directly.

Keywords: Voicemail, speech access, *What You See Is Almost What You Hear*, asynchronous communication, "speech as data", empirical evaluation.

INTRODUCTION

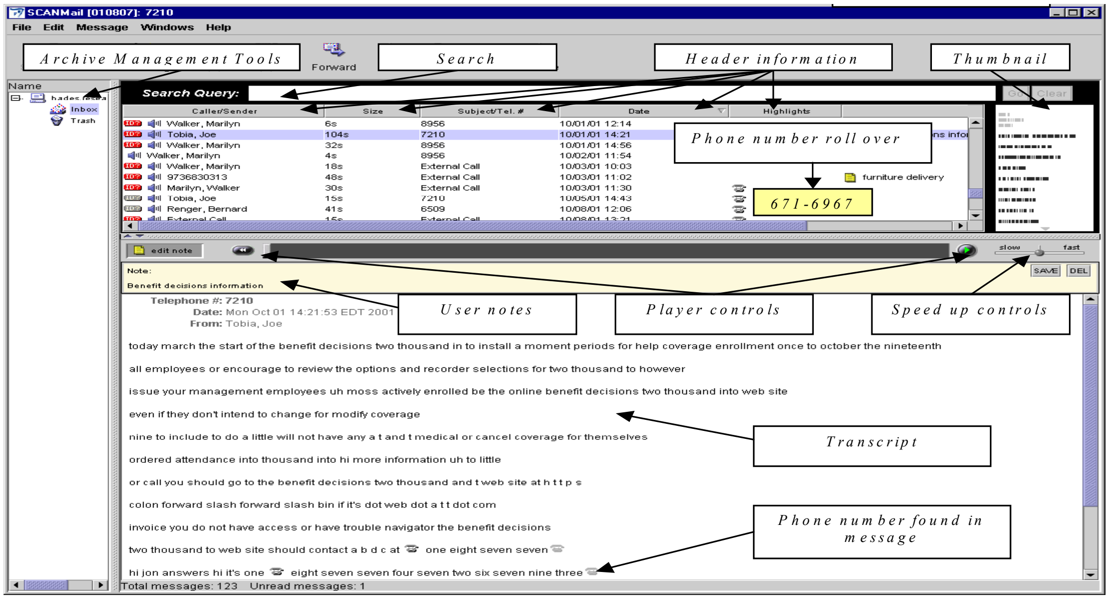
There are increasing amounts of on-line public, corporate, and private speech data, including broadcast news, corporate announcements, meeting records and voicemail archives. Such speech data has general benefits over text, being both expressive and easy to produce [1,4,7,13,20]. Currently, however, its use is hampered by the lack of effective end user tools for accessing and manipulating it. Speech is a serial medium that does not readily support search, visual scanning or key word spotting [7,21]. We can contrast this with tools for accessing text. Text has the benefit of being *searchable* using information retrieval techniques. It also supports *visual scanning*, whereby users navigate to relevant regions using a combination of formatting information and word spotting. They then apply systematic processing to these relevant regions [2]. The goal of the current research is to provide UIs to support both *search* and *visual scanning* of speech data.

Previous approaches to speech access have employed three main techniques. The first uses different types of *structural indices* to access different speech regions: by speaker [5,7,11,24], emphasis [1,17], external events such as user note-taking behaviors [7,12,17,18,22], or accompanying visual events [3,6,11]. Indices are then represented visually, allowing browsing and random access to relevant regions. The second approach involves *content-based search.* Automatic speech recognition (ASR) is applied to speech, and the resulting errorful text is then searched using information retrieval techniques [10,11]. The third is *surface manipulation.* Signal processing techniques are applied to digital speech allowing it to be played back at several times its normal rate, retaining comprehensibility [1]. Our own research combines all three techniques in order to access speech data.

We chose voicemail as our reference application for several reasons. Voicemail is a pervasive, but with a few notable exceptions [13,14], little studied, workplace communication technology, with an estimated 68 million users worldwide [18]. Many organizations rely heavily on voicemail for conducting everyday work, and voicemail is often preferred to email [13,20]. Voicemail is also a common feature of most new cellular phones. Yet despite its ubiquity and importance, there are still many problems with current voicemail user interfaces.

The structure of the paper is the following. We summarize three previous studies of voicemail usage [18,20,21], which served to motivate our design. Those studies identified four key user problems: *message scanning, information extraction, status tracking* and *archiving.* We also identified a central user strategy for voicemail processing, *message transcription*, in which users transcribe all or part of a message in order to avoid re-accessing the original message. We describe SCANMail, a voicemail UI that embodies the principle of *What You See Is Almost What You Hear (WYSIAWYH).* The system generates transcripts of speech data by applying ASR to the underlying speech data. These transcripts are then used as an interface to that speech data. Information retrieval indexing of the transcript provides search and a graphical user

CHI 2002, April 20-25, 2002, Minneapolis, Minnesota, USA.
Copyright 2002 ACM 1-58113-453-3/02/0004…$5.00.



Caller/Sender	Size	Subject/Tel. #	Date	Highlights
Walker, Marilyn	6s	8956	10/01/01 12:14	
Tobia, Joe	104s	7210	10/01/01 14:21	ns infor
Walker, Marilyn	32s	8956	10/01/01 14:56	
Walker, Marilyn	4s	8956	10/02/01 11:54	
Walker, Marilyn	18s	External Call	10/03/01 10:03	
9736830313	48s	External Call	10/03/01 11:02	furniture delivery
Marilyn, Walker	30s	External Call	10/03/01 11:30	☎
Tobia, Joe	15s	7210	10/05/01 14:43	☎
Renger, Bernard	41s	6509	10/08/01 12:06	☎
External Call	15s	External Call	10/08/01 13:21	☎

Note:
Benefit decisions information

Telephone #: 7210
Date: Mon Oct 01 14:21:53 EDT 2001
From: Tobia, Joe

today march the start of the benefit decisions two thousand in to install a moment periods for help coverage enrollment once to october the nineteenth
all employees or encourage to review the options and recorder selections for two thousand to however
issue your management employees uh moss actively enrolled be the online benefit decisions two thousand into web site
even if they don't intend to change for modify coverage
nine to include to do a little will not have any a t and t medical or cancel coverage for themselves
ordered attendance into thousand into hi more information uh to little
or call you should go to the benefit decisions two thousand and t web site at h t t p s
colon forward slash forward slash bin if it's dot web dot a t t dot com
invoice you do not have access or have trouble navigator the benefit decisions
two thousand to web site should contact a b d c at ☎ one eight seven seven ☎
hi jon answers hi it's one ☎ eight seven seven four seven two six seven nine three ☎

Total messages: 123 Unread messages: 1

Figure 1 - SCANMail user interface

interface supports visual scanning, browsing and annotation of the underlying speech. The system also applies information extraction techniques to identify critical pieces of information, such as telephone numbers, from the speech automatically.

We discuss how SCANMail addresses the four voicemail tasks we identified. We describe a laboratory evaluation and a two-month field trial evaluation of the system. SCANMail outperformed a state of the art voicemail system, Avaya's Audix^{TM}. This was attributable to SCANMail's support for visual scanning, search and information extraction. While ASR transcripts contain errors, they nonetheless provide enough information for users to extract key points or to navigate to important regions of the underlying speech, which they can then play directly. In either case, people avoid listening to the entire message.

VOICEMAIL TASKS AND PROCESSING STRATEGIES

In three previous studies [18,20,21], we collected qualitative and quantitative data to identify users' key tasks and strategies for processing voicemail, for a typical voicemail system, Audix. Our data included server logs, surveys and interviews with high volume users. Our data show that voicemail messages contained significant amounts of information: about half those surveyed reported average message lengths of between 30-60s and about half reported lengths of 1-2 minutes. Our interviews also indicated that voicemail messages contain complex information, not simple *"call me back"* requests: *"[a voicemail message] is really like a whole memo, or a huge email message worth of information."* Furthermore, voicemail often substituted for a series of face-to-face meetings: *"entire transactions or entire tasks are accomplished by exchanging [voicemail] messages. That is, you will never talk to the person in real time."*

Users report four main tasks when processing voicemail: *scanning and searching* the mailbox to identify important messages; *extracting information* from individual messages; *tracking the status* of current messages; and *managing* their archive of stored messages.

<u>Scanning and search</u>: *Scanning and search* are used for *prioritizing* incoming new messages, and for *locating* valuable saved messages. Users' current scanning strategy is to sample all messages in sequence to determine location and status. For prioritization, only 24% of people we surveyed use voicemail message headers to identify urgent messages, reporting they are "too slow". Instead users listen to the first few seconds of each message, to the speaker's intonation, to determine whether a message requires immediate action. In *locating* stored messages, most users do not retain a detailed model of their archive and 76% of those surveyed report that "listening to each message in sequence" is their standard procedure for finding archived messages. However, the linear nature of mailbox search makes location onerous when multiple messages are stored.

<u>Information Extraction:</u> When a relevant message is identified, users have to *extract* critical information from it. This is often a laborious process involving repeatedly listening to the same message for verbatim facts such as caller's name and phone number. Multiple listens are also necessary with vague or highly detailed messages. Of those surveyed, 46% report that they replay messages "about half the time". To reduce repetitive processing, 72% of survey users report "almost always" taking *written notes*. Users employ two different note-taking strategies. The first strategy is *full transcription*: here users attempt to produce a written transcript of the target message, to reduce the need for future access. The second strategy is to take notes as *indices*. According to our users, voicemail messages are structured, and the object of this strategy is to abstract the predictable key points of the message (such as caller name, caller number, reason for calling, important dates/times and action items). In most cases, however, users also keep the original voice message as a backup for these incomplete and sometimes sketchy notes. Notes are then used to identify and navigate

within the original message, although users also commented on the laborious nature of constructing and managing these notes.

Status tracking: Workplace tasks are often delegated through voicemail, and a common user problem is *tracking message status.* Status tracking is a prevalent problem for users accessing voicemail under time pressure. They often defer processing a significant number of incoming messages. When accessing voicemail later, they are often unclear about which messages they have dealt with. There are two main techniques for status tracking. In the first, people use notes taken during *information extraction* as reminders. Notes taken on scraps of paper, or a dedicated logbook in the user's work area, remind them about what needs to be done. With the second status tracking strategy, users take no notes but leave undischarged messages in their voicemail mailbox. Reminding takes place when users next scan their archive. In the course of scanning they are reminded about outstanding undischarged messages. The weakness of this second strategy is that there is no visible reminding cue, so that if users do not access the voicemail archive they are unaware of the presence of unresolved items.

Archiving: Users also have to *manage their archives*. Given their access strategies, most users' archives consist of a backlog of undischarged messages as well as saved valuable messages. They therefore engage in periodic "clean-ups": accessing each message in sequence to determine whether it should be preserved. By removing superfluous messages, users also make it easier both to scan for existing valuable messages, and to monitor reminder messages. Those who do not engage in "clean-ups" report being surprised by the extent to which they accumulate irrelevant messages.

SCANMAIL SYSTEM DESIGN

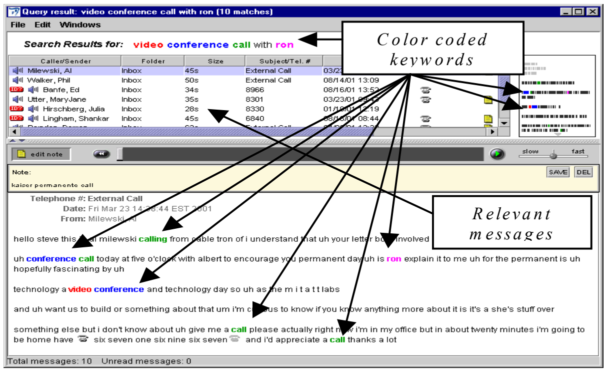
On the basis of these studies we designed a novel multimodal system to support improved voicemail access. We took as a guiding design principle, the user strategy of message transcription, but we also wanted to directly support the users' tasks we had identified. As we have seen, transcribed notes capturing the content of messages (a) assist in information extraction and (b) serve as a *visual analogue* to stored speech messages. Notes also supported scanning, status tracking, and some aspects of archive management. In previous research we showed the utility of a voicemail system that allows users to take manual online notes and access speech using those notes [18]. Despite the benefits of notes, however, users still felt note-taking to be highly laborious.

We therefore designed a system that *automatically* generates transcripts of messages, using ASR. These transcripts are used as the interface to the underlying speech, supporting visual scanning and search. We call this design principle for speech access *What You See Is Almost What You Hear (WYSIAWYH)* [19]. The ASR transcripts usually contain errors because of the current limits of ASR. Nevertheless, they provide enough information for users to extract key points by reading. Or, if the transcript is hard to understand, users can visually scan it to navigate to important regions of the underlying speech. The user interface then allows users to select important regions of the transcript and play the speech corresponding to those specific parts of the message. In both cases, people avoid listening to the entire message. The UI is shown in Fig. 1. There are six main UI elements: headers, transcript, thumbnail, search, player and archive management tools.

The *message header* panel includes: callerID (if available from the phone switch or callerID server), length in seconds, phone number (if available from the phone switch), time and date, any phone numbers extracted from the message, and the first line of any attached user notes. Double clicking on a header selects a message, displays its ASR transcript and begins playing the message.

As we have seen, caller identity information is important for message processing, and one commonly reported limitation of traditional voicemail headers is the absence of callerID information for many calls. For example, for a sample of 3014 messages we logged, the telephone switch only provided callerID information for 36% of messages; external calls, for example, are identified only as such. Even when systems do support external information, this is sometimes not useful. For example, in corporate settings the number provided is often the local switch (PBX) and not the originating phone number. To address this, we built a callerID server that uses user-trained acoustic modeling to automatically identify repeat callers, even when no information is available from the telephone switch. Users provide initial caller labels for messages, along with feedback about whether the callerID system subsequently identified the caller correctly. This feedback information is used to train the callerID acoustic models, and these are currently successful on 91% of messages. Caller names generated by this server are presented as headers when available.

We also use information extraction techniques to automatically identify telephone numbers and names mentioned in the message. 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. We allow users to edit the phone numbers if these are incorrectly transcribed. In a test of 10,000 messages, phone numbers were identified by the program with 94% combined precision and recall.



Search Results for: video conference call with ron

Caller/Sender	Folder	Size	Subject/Tel. #	
Milewski, Al	Inbox	45s	External Call	03/2
Walker, Phil	Inbox	50s	External Call	08/14/01 13:09
ID? Banfe, Ed	Inbox	34s	8966	08/16/01 13:52
Utter, MaryJane	Inbox	35s	8301	03/23/01 09:43
ID? Hirschberg, Julia	Inbox	28s	8330	01/10/01 12:19
ID? Lingham, Shankar	Inbox	45s	6840	08/16/01 08:44

Note:
kaiser permanente call

Telephone #: External Call
Date: Fri Mar 23 14:38:44 EST 2001
From: Milewski, Al

hello steve this is al milewski calling from cable tron of i understand that uh your letter box involved

uh conference call today at five o'clock with albert to encourage you permanent day uh is ron explain it to me uh for the permanent is uh hopefully fascinating by uh

technology a video conference and technology day so uh as the m i t a t t labs

and uh want us to build or something about that um i'm curious to know if you know anything more about it is it's a she's stuff over

something else but i don't know about uh give me a call please actually right now i'm in my office but in about twenty minutes i'm going to be home have ☎ six seven one six nine six seven ☎ and i'd appreciate a call thanks a lot

Total messages: 10 Unread messages: 0

Figure 2 - Search in SCANMail

The *transcript* is generated by an ASR server. Mean transcript word accuracy is currently 76% [8]. Speech is also formatted into "audio paragraphs" using an acoustic segmentation algorithm. Segments are identified using pause duration data, along with information about changes in acoustic signal energy [9]. Clicking on a paragraph or highlighting a specific region of text in the transcript plays the speech corresponding to the region selected. Phone icons also bracket hypothesized numbers in the ASR transcript and clicking on the phone icon plays the speech in which the number was mentioned. Future items to be extracted from messages include dates, and times.

Users also see a *thumbnail* image of the current message depicting a graphical overview of the message. As with the transcript, users can click on the thumbnail to play entire messages or audio paragraphs. They can also associate notes with a message; any note attached to the current message is displayed when the message is selected. A *search* bar allows users to search the contents of their mailboxes (Fig. 2). Search results are presented in a new search window, with keywords color-coded in the query, transcript, and thumbnail. Color-coding depicts regions in the transcript and thumbnail where keywords appear, allowing users to focus on regions in the transcript and thumbnail containing relevant key words. The *player* also supports various audio playing operations, including playing the entire message, or manually selected audio paragraphs. Audio playing speed can also be customized allowing messages to be speeded up or slowed down during playback, while preserving their comprehensibility. *Archive management tools* potentially enable users to file messages into folders or delete them.

Finally we provided an *email server* that forwards ASR transcripts, and message headers to the user's email along with the original voice message stored as an audio attachment (Fig. 3). Users can read the transcript in email or listen to the audio, either by launching the SCANMail client UI or by playing the attachment directly.

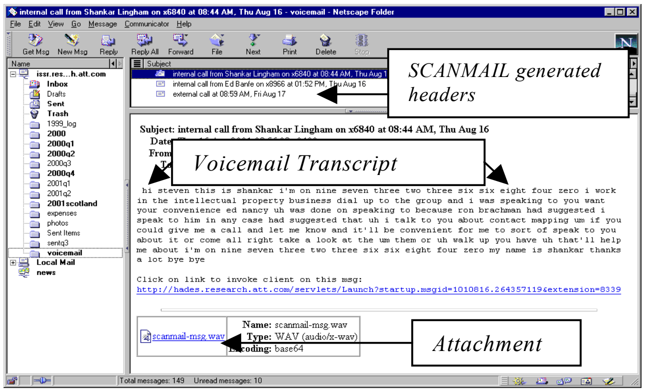
We now describe how the *WYSIAWYH* interface is intended to support the four tasks observed in our original user studies.

Information extraction using message transcripts and headers. In our initial user studies, a key user strategy for *information extraction* was the use of note-taking to textually record significant message content. In SCANMail the message transcript is generated automatically by ASR. Users can extract information from messages in two ways. First, they can read the transcript directly. Second, if the transcript is too hard to understand due to recognition errors, they can use the transcript to navigate to regions they judge important. They can then play the speech for just those regions, by selecting the relevant part of the transcript. We also provide direct assistance for information extraction by automatically extracting key facts such as callerID and phone numbers, and highlighting these in headers and transcripts.

Scanning and search. An important set of cues for prioritizing and locating important messages is provided by header information [18]. By depicting this general information we enable users to visually scan and randomly access messages, so that they no longer have to access messages *in sequence* to identify specific messages. By automatically extracting phone numbers and caller names we facilitate the scanning process. More importantly, we provide *search* that allows users to directly access messages by content.

Status tracking using annotations and overview information. The user interface was designed to support *status tracking* in two ways – again by analogy with users' paper based strategies of leaving themselves visual reminders. User notes can *explicitly* record the actions necessary for each message. More *implicitly,* we hoped that the mere fact of having a visual representation of each message visible in the mailbox would serve to remind people of the necessary action whenever they access SCANMail [23]. For example seeing a message from "Joe Tobia" might remind one of the action that message requires. A final cue to message status is that unaccessed messages are depicted in **bold**, as in standard email clients; once accessed, the font changes.

Archive management. Users reported problems in remembering the contents of their archive, and in preventing the build up of irrelevant messages. The SCANMail interface provided them with a set of tools for managing voicemail. They can create folders, as well as move, copy and delete information in those folders. More implicit support for archive management is provided by the visibility of messages, enabling the archive to be quickly scanned to identify important messages and filter out superfluous ones.


internal call from Shankar Lingham on x6840 at 08:44 AM, Thu Aug 16 - voicemail - Netscape Folder
File Edit View Go Message Communicator Help
Get Msg New Msg Reply Reply All Forward File Next Print Delete Stop
Name
issr.res...h.att.com
Inbox
Drafts
Sent
Trash
1999_log
2000
2000q1
2000q2
2000q3
2000q4
2001q1
2001q2
2001scotland
expenses
photos
Sent Items
sentq3
voicemail
Local Mail
news
Subject
SCANMAIL generated headers
Voicemail Transcript
Attachment
Total messages: 149 Unread messages: 10

internal call from Shankar Lingham on x6840 at 08:44 AM, Thu Aug 1

internal call from Ed Banfe on x8966 at 01:52 PM, Thu Aug 16

external call at 08:59 AM, Fri Aug 17

Subject: internal call from Shankar Lingham on x6840 at 08:44 AM, Thu Aug 16

Date

From

T

```
 hi steven this is shankar i'm on nine seven three two three six six eight four zero i work
in the intellectual property business dial up to the group and i was speaking to you want
your convenience ed nancy uh was done on speaking to because ron brachman had suggested i
speak to him in any case had suggested that uh i talk to you about contact mapping um if you
could give me a call and let me know and it'll be convenient for me to sort of speak to you
about it or come all right take a look at the um them or uh walk up you have uh that'll help
me about i'm on nine seven three two three six six eight four zero my name is shankar thanks
a lot bye bye

Click on link to invoke client on this msg:
http://hades.research.att.com/servlets/Launch?startup.msgid=1010816.264357119&extension=8339
```

scanmail-msg.wav

Name: scanmail-msg.wav
Type: WAV (audio/x-wav)
Encoding: base64

Figure 3 - A SCANMail message forwarded to email

Implementation

Messages are first retrieved from a voicemail server, then processed by the ASR server which segments the speech into "audio paragraphs" and provides a text transcript. Multiple ASR passes are needed to achieve 76% accuracy. To ensure system responsiveness, as each pass is completed it is presented in the UI replacing the previous pass. The first pass is approximately twice real-time and the final pass 12 times real time. The message audio and/or transcript are then passed to the information extraction (IE), information retrieval (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 voicemail messages [8]. Transcripts are indexed by the information retrieval server using the SMART IR [15] engine. Key information, such as phone numbers, is extracted from the transcript by the IE server. The CallerID server builds up a series of acoustic models corresponding to different callers and then hypothesizes caller names for messages for which no caller information is supplied by the switch.

There were two parts to the evaluation: a laboratory study and a field trial. The laboratory study evaluated SCANMail under controlled conditions. The field trial allowed us to examine use of the system for everyday work.

LABORATORY STUDY

In this study, we compared SCANMail with a state of the art voicemail system: Audix. The study had two main goals: to determine (a) whether SCANMail outperformed Audix for experienced users retrieving their voicemail; (b) which system features were responsible for this.

There were 8 users, all of whom had extensive experience with the Audix system (mean 3.8 years). They carried out tasks with both Audix and SCANMail interfaces. We gave them 2 artificial mailboxes each containing 20 messages, and three types of representative task (based on our prior user interviews). The tasks were: local *information extraction* (find a message about a specific topic and extract a specific fact, e.g. a name or number); *search and scanning* (find the message that is most directly related to a given topic), *summary* (summarize a given message with respect to a given question). Summarization can also be viewed as global information extraction. We used artificial mailboxes to control message content, but this meant that we were not able to examine status tracking or archive management tasks, both of which require access to personal data.

Users were encouraged to "think aloud" during the study, so that we could identify their processing problems and strategies. We logged both *process* measures, e.g. number and type of user interface operations, as well as *outcome* measures, task completion times and solution quality (as rated by 3 independent judges). Users differed greatly in their thoroughness: some spent large amounts of time trying to optimize solutions, whereas others were satisfied with quick, but approximate, solutions. As in our previous research [19], to control for this, we used a normalized measure of success, namely *(quality of solution)/(time to solution)*. Users answered a short survey after completing each task, asking them how effective the UI had been for that task. After completing all tasks, users answered a longer survey comparing the two interfaces and evaluating features of both. We also interviewed users after the experiment about their experience with the system.

Our hypotheses and findings were as follows:

H1: *Overall Performance* Predict better performance for SCANMail for normalized solution quality across the 3 tasks because it provides more direct support for all aspects of speech access.

Finding: Data were analyzed using two way analysis of variance, (ANOVA) with task type (information extraction, search, summary) and user interface (SCANMail, Audix) as independent variables and normalized success as dependent variable. As predicted, (see Fig. 4) we found higher normalized scores overall for SCANMail than Audix ($F_{(1,90)}$= 4.02, $p<0.05$).

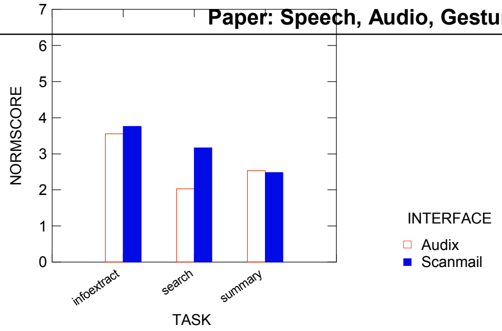


Figure 4 -Normalized performance for two interfaces

H2: *Task-specific effects* Predict greater advantages for SCANMail in local information extraction and search than summary tasks, as both these tasks require cross-message navigation that is not well supported in Audix.

Finding: As predicted, there were task differences ($F_{(2,90)}$= 6.47, $p<0.005$) (see Fig. 4). However planned comparisons showed higher normalized scores for SCANMail for search tasks only (Tukey, $p<0.05$). One reason for the failure to find effects for information extraction is that this task sometimes required access to proper names which are not well recognized by ASR. Users may therefore have to listen in order to extract this name information, reducing the benefit of the SCANMail transcript.

H3: *Overall Preferences*: Predict overall user preferences for SCANMail, because it provides more direct support for all aspects of speech access.

Finding: SCANMail was judged to make tasks 'less time-consuming' ($F_{(1,90)}$= 32.54, $p<0.0001$); to make tasks 'easier' ($F_{(1,90)}$= 8.02, $p<0.001$); to make it 'easier to find relevant messages' ($F_{(1,90)}$= 21.64, $p<0.0001$), and to make it easier to 'find information within messages' ($F_{(1,90)}$= 7.88, $p< 0.01$). (All analyses are ANOVAS, with task and interface as the independent variables and the relevant judgment as dependent variable).

<u>User behaviors</u> We also made predictions about the effects of transcripts and search on performance. We first excluded one user from the analysis because he took detailed notes for every message before responding to any question, and never used search throughout the experiment.

H4: Better quality transcripts should lead people to play less speech, because they can extract more information directly from the transcript or focus playing on important regions of the transcript. This should lead to more efficient speech processing and higher normalized scores.

There was weak evidence for this prediction. Transcripts with lower word error rates were played less ($r_{(41)}$=0.27, $p<0.09$). Furthermore, people who listen to less audio achieved higher normalized scores ($r_{(41)}$=0.40, $p<0.01$), suggesting that good transcript quality improved performance. User comments also indicated that use of the transcript combined with the playbar enabled them to reduce the amount of speech played, either by reading relevant information directly or by using the transcript to navigate to relevant regions for playing.: *"even though the transcripts aren't that precise, they still allowed me to find relevant portions which I listened to using the playbar".*

H5: We also predicted that successful use of search would enable users to play less speech, and perform better. Search should be effective for identifying relevant messages and reduce the potential number of messages that have to be listened to.

Consistent with this, we found that users who successfully generated search queries tended to play less speech ($r_{(41)}$=0.30, p<0.05) and achieve higher normalized scores $F_{(1,40)}$=3.72, p=0.06). User comments also provided support for the utility of search: *"search allowed me to find relevant messages or at least filter to a subset which I then looked at manually."*

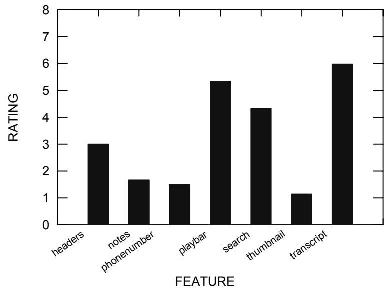


Figure 5 - Perceived feature utility for SCANMail

Despite these overall objective advantages for transcript and search, there were, nevertheless, some disadvantages to the SCANMail interface. Errors in transcription meant that reading the transcript could be misleading or queries could fail. Users also commented that transcripts were hard to read not only because of ASR errors, but also lack of punctuation and formatting. Users who spent large amounts of time attempting to interpret poor transcripts (as indicated by large amounts of scrolling) tended to score poorly ($r_{(41)}$=-0.45, p<0.005).

Some users failed to consider the possibility that words may have been misrecognized by ASR, so that key terms may be missing from the transcript and hence relevant documents not found by search. Other users were aware of the problems with transcript quality and search: *"search is good for finding relevant messages but it's easy to forget that it depends on the quality of speech recognition."* The user who did not use search stated: *"I don't trust search because I might wind up down a blind alley."*

Perceived utility of system features. Users were asked to rate system features from 1 ('of little use') to 7 ('extremely useful'). An ANOVA of perceived feature utility by task showed major differences between features ($F_{(6,315)}$= 8.02, p <0.0001), with feature utility depending on task type ($F_{(12,315)}$= 5.73, p<0.0001). Fig. 5 shows that the search, transcript, and playbar were the most highly rated features, whereas headers, notes, thumbnail, phone number and were not found to be as useful, observations that are supported by post hoc Bonferroni tests. Nevertheless, users pointed out real-life scenarios in which the lower rated features would have been useful, e.g. notes for when they had to repeatedly access an important message, headers and phone number when they were anticipating a call from a specific person.

Overall, the laboratory study showed that SCANMail outperformed Audix, even with highly experienced Audix users. More importantly, it revealed which aspects of SCANMail led to these benefits. Behavioral data showed that despite transcript errors, successful SCANMail users played less audio, because the transcript enabled them to read information directly or to focus on relevant regions of the message. SCANMail search also allowed more efficient audio processing, enabling users to focus on relevant messages. The utility of search and transcript is further evidenced by user preferences for these features when combined with the playbar.

FIELD TRIAL

We designed an 18-user field trial to investigate how effectively SCANMail supported everyday voicemail access. We logged usage data for 8 weeks. We collected data about: the number and duration of SCANMail sessions, messages stored and accessed, operations on messages, notes taken, and the use of search and header data. We also collected survey data comparing SCANMail and Audix. Space constraints prevent a complete presentation of this data. Our main focus is therefore on the user logs, and the evidence these provide about the utility of SCANMail features. We present data for a total of 1746 messages, in 230 user sessions.

One central question concerns the utility of the transcripts for information extraction. Log analysis shows that on 24% of occasions that users accessed a voicemail transcript using SCANMail, they did not play it at all. On these 24% occasions, it seems that users were able to extract sufficient information from the message by reading it, without needing to play it. User comments also showed that people were able to extract the gist of messages without listening to them: "*[SCANMail] lets you see unheard messages at a glance. It's possible to get a sense of a message's topic by glancing over it."* Users also pointed out that many messages had prior context. This meant that even when transcripts contained errors, users could still make sense of them. Furthermore, when users did play messages, on 52% of occasions they did not play the complete message. In addition, 57% of SCANMail play operations were originated by clicking on parts of the transcript. This partial playing of messages also indicates the utility of the transcript in allowing users to restrict their audio processing to important regions of the message. We can contrast the greater efficiency of SCANMail with Audix processing, where users generally process messages by listening to them in their entirety.

Another source of evidence for the utility of transcripts is users' note-taking behaviors. Logs show that users took on average only 2.61 notes, i.e. for only 4% of messages accessed through SCANMail. This again contrasts with Audix processing where note-taking is the rule. The decreased amount of note-taking with SCANMail seems to occur because the transcript obviated the need for large numbers of manual notes: *"I previously took paper notes to record name, phone number and a synopsis of the message. With SCANMail that information is so easy to retrieve that I don't need notes."* Users also liked the fact that the system generated this information for them automatically, as opposed to having to take manual notes: *"I like having a record of the transcript typed for me."*

While transcripts were highly valuable for information extraction, other system features also seemed to prove useful. Extracted phone numbers were viewed from headers on 37% of occasions that a message was accessed. Furthermore, on 6% of occasions, accessing a phone number was sufficient for message processing, as users did not read or play any other part of the message. The accuracy of the extracted phone numbers is also shown by the fact that only one user corrected a phone number.

The data concerning the utility of search are more mixed. Log analysis indicates that users retained **98%** of all messages they received. The increased numbers of archived messages suggest that, in contrast to Audix, users have shifted from regarding voicemail as an ephemeral set of messages that have to be processed immediately, to viewing voicemail as a more permanent informational resource. One user commented: *"(SCANMail) provides a permanent visible record of all my voicemail messages."* While these data show the shift to an archival view of voicemail, search itself was only used on 1% of occasions that messages were accessed, even though search was perceived overall to be 'somewhat useful' in the user survey. Search may not be used much, however, because people's archives were relatively small, allowing them to access messages by visual scanning.

Email forwarding and unanticipated uses of SCANMail

Users were extremely positive about having voicemail transcripts and messages forwarded to email. Email forwarding served to notify them about the presence of incoming voicemail. Furthermore, providing email transcripts allowed users to determine the key points and importance of the message, and only if necessary play the relevant portions. Many users pointed out the efficiency of accessing messages directly through email compared with the complexity of logging into the voicemail system. Greater ease of access may also increase responsiveness to voicemail: *"since I check email more frequently than voicemail I responded more promptly to voicemail."* Several people thought voicemail forwarding provided a unified messaging interface to both email and voicemail: *"On weekends when I checked my email I was also checking my voicemail, that was way cool and extraordinarily useful."* SCANMail also led to more call screening. One user pointed out that the ease of accessing and processing messages with SCANMail meant that he altered his handling of incoming calls. With SCANMail he was more likely to screen calls by letting them go through to voicemail. Such screening was particular popular with cellphone users, whose phones included a small display. These users would forward their email to their cellphones and check the first few lines of voicemail transcriptions, to decide whether it was important to access the full message.

CONCLUSIONS

Based on data gathered from user studies, we built a novel UI to speech data based around ASR transcripts that are used to allow visual scanning, information extraction and search. The voicemail UI supported user tasks, both in field trial and experimental settings, outperforming a state of the art voicemail system. Our analyses show that providing transcripts facilitated more efficient message processing, either by enabling users to read information from the transcript directly or using the transcript information to focus their audio playing on important regions of the message. Automatic extraction of telephone numbers was also valuable. The field trial also indicated a shift from users viewing voicemail as an *ephemeral medium* requiring immediate processing to an *informational archive*. Most users quickly habituated to the idea of message permanence and archiving. Although search was infrequent, information extraction tools seemed to convert voicemail into a viable informational resource, in contrast to current touchtone systems that make search and scanning highly onerous. An email forwarding service also proved extremely successful in providing rapid scanning and efficient access to voicemail through email transcripts and speech attachments. Together these findings vindicate the *What You See Is Almost What You Hear* principle for speech UI design in which the transcript is viewed as a *visual analogue* to the underlying speech, and used as an interface for accessing speech data.

This demonstration of transcript utility for speech browsing and search is consistent with our other work based on access to broadcast news data [19]. Our WYSIAWYH approach is different from other systems designed to access large speech corpora that do not use the transcript for access [3,6,10,11]. In these systems, speech is accessed using indices extracted from video or by semantic processing. How can we explain our greater success in using the transcript? One critical difference between our work and [3,6,10,11] lies in the transcript representation. In our interface, the transcript is formatted (using acoustic segmentation), and active, allowing users to browse to important regions, select and play the underlying speech. These other systems do not allow the transcript to be used to control playing in this way. Another factor is ASR accuracy. Elsewhere we have shown that accuracy has a large effect on the success of speech browsing and search [16]. It may be that in these other systems, transcript accuracy was not sufficient to allow information extraction or directed browsing.

Nevertheless, there are some disadvantages to the approach. Some of these stem from ASR errors. While the transcripts are generally useful, placing too much trust in inaccurate transcripts can induce errors in information extraction and search. One way to address this in future UIs, might be to depict ASR confidence information, i.e. information about the probability that a given word was correctly recognized. Thus words for which the ASR had lower confidence might be grayed out to suggest they should be treated with caution. Another possibility suggested by several users is that transcripts are editable, allowing errors to be corrected. Such a system might combine the advantages of an annotation based voicemail system [18], while avoiding the onerous task of taking handwritten notes.

Users also pointed out that the current system is desktop-based, whereas much of their communication work requires mobile access to voicemail. Some users took to forwarding their email (including voicemail transcripts) to their cellphones to address this. We have therefore recently developed new mobile versions of SCANMail with simplified UIs that run on the Compaq iPaq and cellphones. Users also reported in the field trial that the UI seemed to be ineffective in supporting status tracking tasks. One explanation is that the prior lack of mobile access to SCANMail means that users are not reminded about outstanding voicemails while away from their desks. Again our new mobile systems should address this.

Finally, there are both practical and theoretical implications to our results. Our tool successfully addresses a significant problem for many users - namely *efficient* voicemail access. Our data also contribute to a growing body of research on general methods for speech access [1,5,7,10,11,18,19,22]. We have also demonstrated the viability of a novel technique of speech access, *WYSIAWYH,* where a transcript provides a visual analogue to underlying speech, supporting scanning and search. Other questions we are currently investigating include: How good does ASR have to be to support scanning and search [16]? Can we extend information extraction techniques to find other important information such as dates, times and locations? Can we determine which messages or parts of messages are important and which are trivial using prosodic and lexical information? Finally we are examining whether our approach be extended to other applications where speech access is important, e.g. meeting capture, focus groups or legal interviews.

REFERENCES

1. Arons, B. Interactively skimming speech. Unpublished PhD thesis, MIT Media Lab, 1994.
2. Askwall, S. Computer supported reading vs reading text on paper: a comparison of two reading situations, *International Journal of Man Machine Studies,* **22,** 425-439, 1985.
3. Boreczky, J., Girgensohn, A., Golovchinsky, G., and Uchihashi, S. An Interactive Comic Book Presentation for Exploring Video. In *CHI2000*, 185-192, 2000.
4. Chalfonte, B., Fish, R., and Kraut, R. Expressive richness. In *CHI91*, 21-26, 1991.
5. Degen, L., Mander, R., and Salomon, G. Working with audio. In *CHI92*, 413-418, 1992.
6. Hauptmann and Witbrock, M. Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval, In M. Maybury (Ed.), *Intelligent Multimedia Information Retrieval,* AAAI Press, pp. 213-239, 1997.
7. Hindus, D., Schmandt, C., and Horner, C. Capturing, structuring and representing ubiquitous audio. *ACM Transactions on Information Systems*, 11, 1993.
8. Hirschberg, J., Bacchiani, M., Hindle, D., Isenhour, P., Rosenberg, A., Stark, L., Stead, L., Zamchick, G., and Whittaker, S. SCANMail: Browsing and Searching Speech Data by Content, *Proceedings of Eurospeech 2001*, Aalborg, 2001.
9. Hirschberg, J. and Nakatani, C. Acoustic indicators of topic segmentation. In *ICSLP98*, 1998.
10. Jones, G., Foote, J., Spärck Jones, K., and Young, S. Retrieving Spoken Documents by Combining Multiple Index Sources, In *SIGIR96*, 30-38, 1996.
11. Kazman, R., Al-Halimi, R., Hunt, W., and Mantei, M. Four paradigms for indexing videoconferences. In *IEEE Multimedia*, 3(1), 63-73, 1996.
12. Moran, T., Palen, L., Harrison, S., Chiu, P., Kimber, D., Minneman, S., van Melle, W., and Zellweger, P. "I'll get that off the audio": salvaging in a multimedia meeting. In *CHI97*, 202-209, 1997.
13. Rice R. and Shook, D. Voice messaging coordination and communication. In C. Egido, J. Galegher and R. Kraut, eds., *Intellectual Teamwork*, Lawrence Erlbaum, NJ, 1990.
14. Rice, R.E., & Tyler, J. (1995). Individual and organizational influences on voicemail use and evaluation. *Behaviour and Information Technology*, *14*(6), 329-341.
15. Salton, G. *The SMART Retrieval System*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
16. Stark, L., Whittaker, S., and Hirschberg, J. ASR satisficing: the effects of ASR accuracy on speech retrieval. In Proceedings of International Conference on Spoken Language Processing, 2000.
17. Stifelman, L., Arons, B., and Schmandt, C. The audio notebook: paper and pen interaction with structured speech. In *CHI2001,* 182-189, 2001.
18. Whittaker, S., Davis, R., Hirschberg, J., and Muller, U. Jotmail: a voicemail interface that enables you to see what was said. In *CHI2000*, 89-96, 2000.
19. Whittaker, S., Hirschberg, J., Choi, J., Hindle, D., Pereira, F., and Singhal, A. SCAN: designing and evaluating user interfaces to support retrieval from speech archives. In *SIGIR99*, 26-33, 1999.
20. Whittaker, S., Hirschberg, J. and Nakatani, C. All talk and all action. In *CHI98*, 249-250,1998.
21. Whittaker, S., Hirschberg, J. and Nakatani, C. Play it again: a study of the factors underlying speech browsing behaviour. In *CHI98*, 247-248,1998.
22. Whittaker, S., Hyland, P, and Wiley, M. Filochat: handwritten notes provide access to recorded conversations. In *CHI94*, 271-277, 1994.
23. Whittaker, S. and Sidner, C. Email overload: exploring personal information management of email. In *CHI'96* 276-283, 1996.
24. Wilcox, L. Chen, F., Kimber, D. and Balasubramanian, V. Segmentation of Speech Using Speaker Identification. *Proc. ICASSP*, 1994.