

Error Correction of Voicemail Transcripts in SCANMail

Moira Burke

Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213
moira@cmu.edu

Brian Amento and Philip Isenhour

AT&T Labs-Research
180 Park Avenue
Florham Park, NJ 07932
{brian, isenhour}@research.att.com

ABSTRACT

Despite its widespread use, voicemail presents numerous usability challenges: People must listen to messages in their entirety, they cannot search by keywords, and audio files do not naturally support visual skimming. SCANMail overcomes these flaws by automatically generating text transcripts of voicemail messages and presenting them in an email-like interface. Transcripts facilitate quick browsing and permanent archive. However, errors from the automatic speech recognition (ASR) hinder the usefulness of the transcripts. The work presented here specifically addresses these problems by evaluating user-initiated error correction of transcripts. User studies of two editor interfaces—a grammar-assisted menu and simple replacement by typing—reveal reduced audio playback times and an emphasis on editing *important* words with the menu, suggesting its value in mobile environments where limited input capabilities are the norm and user privacy is essential. The study also adds to the scarce body of work on ASR confidence shading, suggesting that shading may be more helpful than previously reported.

Author Keywords

Voicemail, error correction, speech recognition, editor interfaces, confidence shading.

ACM Classification Keywords

H.5.2. User Interfaces – evaluation/methodology, interaction styles. I.2.7 Natural Language Processing – speech recognition and synthesis.

INTRODUCTION

Voicemail is a heavily used communication tool and yet is inherently difficult for humans to process, especially in a mobile setting. Unlike text, audio is a serial medium that does not naturally support search or visual scanning.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee

CHI 2006, April 22-27, 2006, Montréal, Québec, Canada.
Copyright 2006 ACM 1-59593-178-3/06/0004...\$5.00.

People must listen to voicemail messages in their entirety, making it difficult to access relevant passages. Interaction with current voicemail systems is cumbersome and users tend to limit their exposure by taking extensive notes and deleting messages immediately. Given the choice, people would like to archive voice messages, provided the proper tools were available for effortless retrieval and review [22]. Retrieval of archived messages, however, depends upon accurate transcription of important words.

More than ever, users require mobile access to their communications applications. While email and chat applications integrated into mobile devices allow discreet access in public situations, most voicemail applications require audio playback, which is both conspicuous and processor-intensive. Our goal is to provide a system for accessing voicemail that improves user performance in the office and simplifies interactions on often underpowered mobile devices.

Numerous tools improve access to speech data. Browsers that use structural indices facilitate navigation to specific regions based on speaker [5,10,12,24], emphasis [16], external text notes [8,13,16,21,23], or accompanying video [3,8,12]. Signal processing techniques maintain comprehensibility in speech played at several times its normal rate [1], and automatic speech recognition (ASR) engines generate text transcriptions of audio content.

SCANMail combines many of these techniques to provide a more email-like interface to voicemail, improving users' access to voice messages [1,22]. SCANMail uses text transcriptions as random access indices into the original audio, giving users more freedom to play back only relevant audio clips. Other important features include keyword search, information extraction, and speaker identification. Technical details of SCANMail appear in the next section.

Despite its benefit over traditional voicemail applications, SCANMail is hindered by the state of speech recognition technology. Previous research has shown that plain text transcripts need to be at least 84% accurate to be comprehensible to readers, but can be below 69% accurate if presented inside a user interface, like SCANMail, that links transcriptions with audio [15]. SCANMail currently generates transcripts that are 82.4% accurate, falling well within the above range, but transcripts may still contain

significant errors. Though users report being able to understand the gist of the message [22], inaccurate transcripts can sometimes be difficult to understand and limit keyword searching. Over-dependence on seemingly correct transcripts may lead users to miss important details. While this is not a problem for short, temporally transient messages that users delete quickly after absorbing, it can be an obstacle for users that wish to archive important messages for later referral.

SCANMail currently addresses these issues in two ways: (1) ASR confidence shading and (2) a transcript editor. With confidence shading, words with a low probability of correct recognition appear in a lighter shade of gray than words of higher confidence, indicating regions that users should treat with caution. Confidence shading, however, does not address misrecognized keywords and phrases that would be important for later message retrieval.

A built-in transcript editor thus allows for quick word- and phrase-level changes, improving comprehensibility without requiring users to take detailed notes separately. Misinterpreted but critical words can be manually corrected for keyword indexing. These corrections are then used to automatically augment the vocabulary, acoustic and language models of our speech recognition system for further retraining. This helps to improve recognition in future messages and increase domain-specific vocabulary without requiring time-consuming, manual modifications of the underlying models. User-contributed corrections are especially beneficial to a system trained on such a general corpus as voicemail. We are in the process of evaluating the effectiveness of such user-contributed corrections.

This paper describes the SCANMail system and an empirical laboratory evaluation of two versions of an embedded transcript editor, one menu-based and one keyboard-based. The former would be useful on platforms without a full keyboard or for novice typists, while the latter would support easy changes to longer phrases, such as when ASR accuracy is low.

SCANMAIL SYSTEM AND EVALUATION

SCANMail presents voicemail messages in a familiar, email-like display (see Figure 1). ASR transcripts of the messages are displayed with playback controls for the original audio. The transcript serves as a visual analogue to the underlying audio data, allowing for easy message browsing. A full description of SCANMail functionality can be found in our previous work [22].

Previous user studies have shown SCANMail to outperform a state of the art voicemail system (Audix) in both task speed and user preference measures. In one lab study [22], eight users with extensive Audix experience (mean 3.8 years) performed three kinds of tasks—information extraction, search and scanning, and summarizing—across 20 messages in both Audix and SCANMail. SCANMail received higher scores overall, with greatest advantage in search tasks, which required cross-message navigation, a strategy not well supported by traditional voicemail applications. Participants strongly preferred SCANMail to Audix, reporting SCANMail to be less time-consuming and easier for finding individual messages and information within messages.

An eight-week field study [22] revealed that SCANMail significantly changed users' voicemail processing

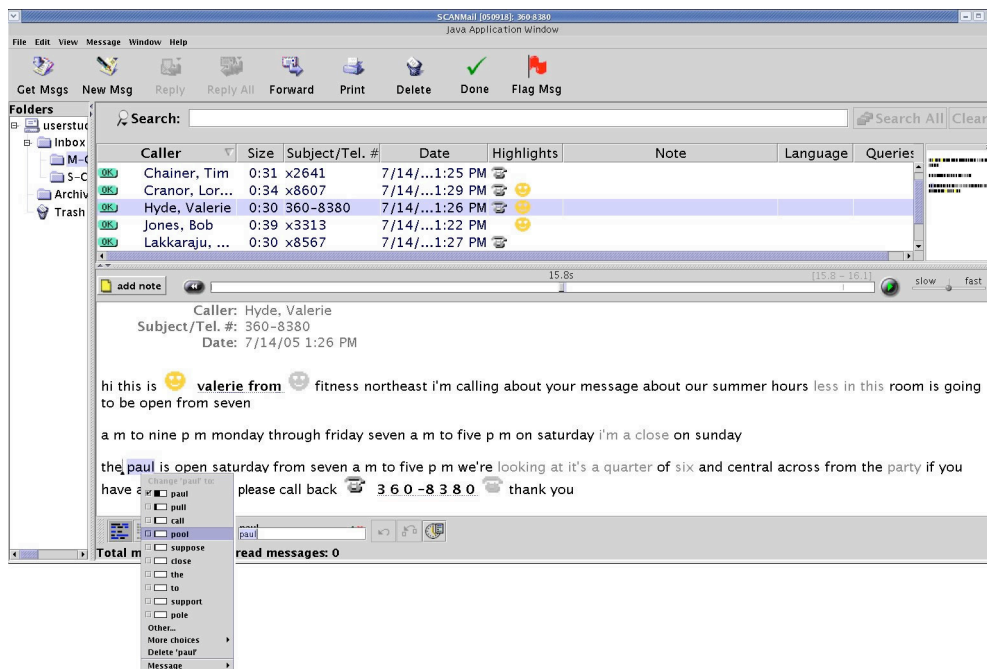


Figure 1. SCANMail interface with an incorrect word highlighted.

behaviors. Participants relied solely on the text transcripts for 24% of the messages—never playing any audio—and did not play the entire audio for 54% of the remaining messages. Users stated that archiving was unimportant, based on past difficulties with their current voicemail archives. With SCANMail, however, they retained 98% of their messages, shifting from the traditional voicemail mindset of quick, ephemeral message access to treating voicemail as a permanent resource.

For voicemail archives to be as useful as email archives, the ability to correct text transcriptions for archival is vital. An eight-month field trial of SCANMail in a network operations center showed that archiving was an important aspect of the deployed system [20]. Without embedded tools for correction, users often added manual annotations of corrected keywords to the voicemail messages to assist with future searches of the archive. Our current system incorporates this feature, simplifying the editing process and reducing the need for external annotations.

Field trials also revealed that users appreciated SCANMail's email forwarding functionality, in which transcripts and headers were sent directly to users' email addresses with the original audio attached. Email forwarding notified them of the presence of incoming voicemail and allowed for simple call screening without the need to log in to a cumbersome voicemail system. Furthermore, it provided a unified interface for voicemail and email messages; messages in both media could be accessed at the same time and archived together: *"On weekends when I checked my email I was also checking my voicemail, that was way cool and extraordinarily useful."* Approximately 100 employees of a large corporate research lab currently use SCANMail, and most have the text transcripts sent directly to their regular email clients. In an informal survey we learned that 71% of these users regularly forward or reply to SCANMail transcripts through email rather than a return phone call and 75% of the time they make manual corrections to the transcript text before sending.

SCANMail runs on numerous platforms, including the Compaq iPaq, Blackberry, and SideKick. A Flash version allows lightweight versions of the playback controls from the full SCANMail application to be embedded in messages forwarded directly to email.

ERROR IDENTIFICATION AND CORRECTION

Identifying and correcting errors in transcripts remains a difficult problem in speech-to-text systems, despite advances in speech recognition technology. Users of commercial dictation systems spend a majority of their time locating and fixing mistakes [11]. Error correction within voicemail transcripts is particularly difficult because, unlike dictation from a single, known speaker, voicemail comes from many different people—including strangers—speaking about a wide range of topics, and the listener may not understand certain words.

Confidence shading is one approach to aid users in finding potentially incorrect words within a transcript [6, 17, 18]. The speech recognition engine assigns a confidence score to each word, and words below a given threshold are displayed in a lighter shade. Confidence scores themselves may be incorrect, failing to detect misrecognized words or falsely catching correct words. Little empirical support for the usability of confidence shading exists because the majority of previous research has focused on using confidence probabilities to improve system recognition accuracy [4,9]. In one notable exception, Suhm and colleagues found that confidence shading within a dictation system did not reduce users' times to locate errors [17]. However, that system treated missed detections and false alarms equally and had a confidence threshold of 0.6 to minimize both kinds of errors. Thus, many misrecognized words appeared to be correct. SCANMail treats missed detections as more serious than false alarms, and so its threshold is set to 0.93 to minimize missed detections. With a threshold of 0.93, SCANMail correctly classifies 88.7% of words, with a missed detection rate of 3.4% for the messages in the current experiment.

Other research investigating the utility of confidence scores shows promising trends but has been inconclusive. To assess user comprehension of recorded speech, Vemuri and colleagues explored the combination of time-compressed audio with speech transcriptions during question-answering tasks. The highest level of comprehension was observed using perfect transcriptions, but this was followed closely by conditions using error-filled, automatically generated transcriptions. One of the latter conditions displayed transcripts with varying word brightness levels, based on a phrase score provided by the recognition engine. Results indicated that the brightness levels improved performance, but the differences over a plain text display were not confirmed to be significant [18].

User-specified error correction methods vary across systems. Typical options include (a) respeaking, (b) typing, (c) choosing from a menu with mouse, voice, or pen, or (d) a combination of the above. Multimodal input generally outperforms simple respeaking [12, 17]. Though humans naturally re-say a word that has been misunderstood in face-to-face dialog, respeaking to a computer often repeats the initial recognition error. Additionally, humans tend to hyperarticulate words when verbally correcting errors, producing speech patterns significantly different from system training data. Commercial dictation systems like Dragon NaturallySpeaking and IBM ViaVoice use voice selection from a menu of alternatives. Users say a command like *"Correct <word>,"* which presents a menu, from which they select a word by number: *"Pick n."* Commercial systems also support replacement by keyboard and mouse selection. Fast typists make repairs most quickly with standard keyboard and mouse, but slower typists and users of mobile platforms may perform well

using a combination of speech and pen input if baseline speech recognition accuracy improves [17].

The present study extends previous research in error correction by investigating voicemail, a system for which users have radically different needs than in the dictation systems commonly studied. Dictation requires long passages of text to be corrected verbatim, while voicemail messages are brief, and people only need to correct sporadic keywords to understand the gist. Confidence shading and error-correction techniques investigated for dictation may have significantly different levels of usability in voicemail, and yet voicemail is arguably more prevalent than dictation applications. Understanding the nature of ASR errors within voicemail and how users correct those errors will greatly improve a daily communication tool, leading to better integration with email and better information retrieval on mobile devices.

ERROR CORRECTION IN SCANMAIL

SCANMail supports error correction in two ways: (1) A menu from which users select replacements at the word level from a list of alternatives, and (2) “swipe-and-type,” in which users select a phrase and type replacement text (see Figures 2 and 3). The menu behaves like a typical spell-checker: Right-clicking a word reveals a list of ten words plus “delete” and “other,” allowing users to type a word not on the list. A “more choices” option reveals a fly-out menu with additional words from the ASR engine, though users rarely opened it more than a few times and said searching the secondary menu took more effort than selecting “other” and typing in a replacement.

The alternative words that populate the menus are generated from a Word Confusion Network (WCN). WCNs are obtained from word graphs, or lattices, that are used to

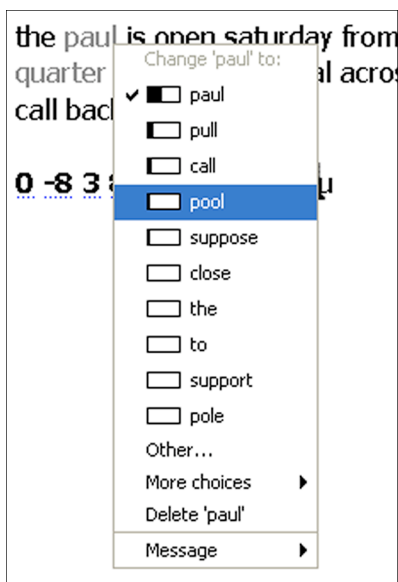


Figure 2. Menu interface for error correction.

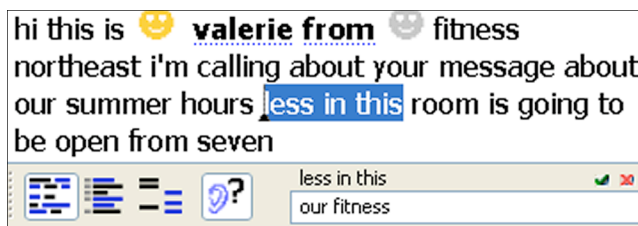


Figure 3. "Swipe-and-type" interface for error correction.

approximate the word search space in large vocabulary continuous speech recognition (LVCSR) systems. Most ASR systems present a one-best hypothesis representing the single best path through the lattice, but there are many additional alternative paths that are ignored. Lattices contain the complete picture of the ASR output but can be unwieldy. WCNs attempt to compress lattices down to a more basic structure that still provides n-best hypotheses for an audio segment. Figure 4 shows the general structure of lattices and WCNs. Competing words in the same approximate time interval of the lattice are forced into the same group in a WCN, keeping an accurate time alignment. Each alternate word in a WCN has a posterior probability—the sum of the probabilities of all paths that contain that word at that approximate time frame—that can be used as a word confidence score [7].

Each editor presents a tradeoff. The menu provides assisted editing, but only of a single word at a time. The swipe-and-type editor allows the user to correct entire phrases in one action, but does not suggest alternative words.

EXPERIMENT COMPARING SWIPE-AND-TYPE AND MENU EDITORS

In this study, participants corrected transcripts of voicemail messages using two SCANMail editors: (1) Menu and (2) swipe-and-type. Messages fell into two categories: (1) High ASR quality and (2) low ASR quality. We measured task performance and user preference to determine the conditions under which each editor would be most appropriate.

Participants

Sixteen participants (ten males and six females) with a mean age of 39 volunteered for a small food reward. All were researchers or administrative staff at a large corporate research lab, representing the expected near-term demographic of SCANMail users. All had five or more years of experience with voicemail and email.

Procedure

Prior to coming to the study, participants filled out a short web survey about their email and voicemail habits. At the study, they began with a brief typing test in order to evaluate their typing speeds. Then they practiced using all of the audio controls on a test message playing within SCANMail until they demonstrated proficiency playing, rewinding, and

selecting regions of messages. Mean practice time was four minutes.

Participants then read brief directions about transcript errors. They were told to imagine they had received ten voicemail messages from friends and coworkers that had information they would want to keep and retrieve again in a few days or weeks. They were instructed to find *important* details with mistakes and correct them; they would not need to correct every error, only those that they deemed important for understanding the message in the future. Thus, the number and quality of corrections were left to the discretion of the participant. They were told to assume proper names were already correct.

Then they used one of the two editor interfaces: menu or swipe-and-type. They corrected errors in a practice message until they were comfortable using the editor. At that point, they were instructed that their edit time for the next four messages would be recorded and they should work quickly but accurately for the corrections they chose to make. Messages appeared one at a time and participants verbally signaled that they had finished editing each one. The experimenter controlled a software timer on a separate computer that recorded mouse and keyboard behavior, revisions, and edit time. After finishing the fourth timed message, they practiced using the second editor and edited four more messages. Messages and editors were counterbalanced across participants; practice messages remained the same for all participants.

After using both editors, participants completed a brief web survey asking them to report the ease and speed of each editor on a series of 5-point Likert scales. The experiment took less than an hour.

Materials

Ten voicemail messages with a mean length of 32 seconds (range 25-44s) appeared within SCANMail. Five came from a naturally occurring corpus from the Linguistic Data Consortium and the experimenters recorded five additional messages. Messages contained general information such as restaurant recommendations, meeting times, and equipment requests. They could be easily understood by participants and did not require additional context for their interpretation.

The following is a representative message:

“Hi, this is Valerie from Fitness Northeast. I'm calling about your message about our summer hours. Our fitness room is going to be open from 7:00am to 9:00pm, Monday through Friday, 7:00 am to 5:00pm on Saturday, and we're closed on Sunday. The pool is open Saturday from 7:00am to 5:00pm. We're located at the corner of Sixth and Central across from the park. If you have any questions please call back, 360-8380. Thank you.”

Transcript accuracy clearly influences how well people understand messages and potentially their ease of editing, so messages of both high and low speech recognition quality were used. To determine quality, we manually created perfect transcriptions for all ten messages and compared them to ASR transcriptions using SCLite, a NIST speech recognition scoring tool. High and low quality transcripts were 81.2% and 37.3% accurate, respectively; these values were one standard deviation above and below the mean for the voicemail corpus.

High quality ASR

SCANMail uses a proprietary internal recognition engine based on Hidden Markov Models. Punctuation and capitalization are not included. Four high quality transcripts were created by running the audio through SCANMail and manually correcting proper names, giving a mean accuracy of 81.2%. The two practice messages were also high quality.

The following is the high quality transcript of the previous message:

“hi this is valerie from fitness northeast i'm calling about your message about our summer hours less in this room is going to be open from seven a m to nine p m monday through friday seven a m to five p m on saturday i'm a close on sunday the paul is open saturday from seven a m to five p m we're looking at it's a quarter of six and central across from the party if you have any questions please call back three six zero eight three eight zero thank you”

Low quality ASR

The four low quality transcripts came from the initial pass of our ASR engine, which uses acoustic feature extraction to rapidly generate a transcript with a mean accuracy of 37.3%.

The following is a representative low quality transcript:

“hey candy it's me hi know about the middle eastern is there had at night after and lab and somebody recommended the i left you can say greenwich the lynch the address is one seventy macdougall it kind of your sixth avenue in bleecker uh near washington square park which is and i think arts is it's probably about forty five minute walk and agent let me know if that's something you bye”

A perfect transcript of the message follows:

“Hey honey, it's me. How do you feel about Middle Eastern food for dinner on Saturday night? I asked around in the lab and somebody recommended the Olive Tree Cafe in Greenwich Village. The address is 117 MacDougal. It's kind of near Sixth Avenue and Bleecker, near Washington Square Park, which is where that big arch is. I think it's probably about a

forty-five minute walk from Penn Station. Let me know if that sounds good to you. Bye."

Word confusion networks

High quality transcripts generated by SCANMail also included word confusion networks, the source of the list of alternative words presented in the menu editor. An XML file contains the transcript and WCN. WCNs were generated for the low quality transcripts by running the audio through SCANMail and manually inserting the resultant WCNs into the transcripts.

The first ten words of the WCN for "paul" (actually "pool") in the first message above were "paul," "pull," "call," "pool," "suppose," "close," "the," "to," "support," and "pole."

Ten words appear on the menu based on an analysis of the word confusion networks for 20 randomly-selected voicemail messages from a corpus of 500. Perfect transcripts were generated manually by the experimenters and compared to the ASR output. The twenty messages contained a total of 212 word errors. The correct word was within the WCN 85% of the time, and in 75% of those cases, the word was within the top ten.

Confidence shading

Confidence shading in SCANMail works as follows: Words below the ASR confidence threshold of 93% are printed in gray; all other text is black. The engine used to create low quality transcripts does not generate confidence scores, so confidence shading was manually added to those messages. To maintain consistency in shading across low and high quality transcripts, incorrect words always appeared gray. Most correct words appeared black but a few appeared gray, based on the default output from SCANMail's ASR. Eliminating missed detections increased the accuracy of the default SCANMail behavior by 3.4% but was deemed necessary to maintain consistency across low and high quality transcripts. Editor ease, rather than confidence shading utility, was the primary factor of interest in this experiment.

Variables and Measures

Editor type

We compared the menu and the swipe-and-type editors. Participants used both editors in counterbalanced order.

ASR accuracy

Four high quality transcripts and four low quality transcripts appeared in counterbalanced order across editors. Participants saw all messages.

Measures

To determine the performance of each editor, we measured (a) audio playback time per message, (b) edit time per message, (c) effort, or number of words changed, and (d) changed word importance.

Hypotheses

On messages with few errors, a few sporadic words will require change, so the menu editor will lead to faster task performance, less effort, and increased user satisfaction. On messages with many errors, participants will perform better with swipe-and-type, as less effort will be required to correct contiguous phrases.

RESULTS

Initial Survey

Validating previous research [22], we found that users want to archive and reuse their messages, but instead often delete them within voicemail. When asked how often they access archived messages in email and voicemail, users responded very frequently (4.88) for email and infrequently (2.31) for voicemail (on a 5-point Likert scale ranging from infrequently to frequently). One obvious reason for this is the difficulty in locating voicemail messages. On a 5-point Likert scale from very difficult to very easy, participants found archived voicemail difficult to locate (2.69) and email easy to locate (4.06). By providing a more accurate transcript for voicemail messages—that users can easily correct—we hope to bridge this gap and make voicemail messages as easy to handle as email messages. When asked if voicemail is harder and more time consuming than email, participants agreed with both statements (4.13 for harder, 4.0 for more time consuming).

Quantitative Results

To investigate the effects of ASR quality and editor type, we conducted a number of ANOVAs with editor type, ASR quality, and typing-speed group (defined below) as independent variables. The main dependent variables were (a) audio playback time, (b) task time, (c) edit operations, and (d) changed word importance.

One of the components of voicemail processing that we would like to minimize is audio playback. Examining client logs, we calculated the amount of time spent listening to the audio for each message. Table 1 summarizes the results. Editor type affected the amount of audio played. In the menu condition, participants played far less audio than in the swipe-and-type condition [$F(1,124)=40.09$, $p<0.00001$]. As expected, high quality transcripts required shorter audio play times than low quality transcripts [$F(1,124)=7.6$, $p<0.01$].

		Audio playback time per message (s)		
		Menu	Swipe	Average
ASR Quality	High	17.7	33.1	25.4
	Low	30.4	50.5	40.5
	Average	24.0	41.8	

Table 1: Participants listened to significantly less audio with the menu than with swipe-and-type.

After subtracting audio playback time, editor type also affected edit time. Table 2 presents the results. Participants spent significantly more time editing with the menu ($M=135.86s$) than with swipe-and-type ($M=100.28s$) [$F(1,124)=9.92, p<0.01$]. As expected, ASR quality also affected edit time [$F(1,124)=42.98, p<0.00001$]. There was no interaction between editor type and ASR quality for edit time.

The reduced audio playback time but increased edit time with the menu resulted in similar overall task times across the two kinds of editors. Participants using the swipe-and-type editor took an average of 150.5s to complete the tasks while the menu condition yielded 154.8s [$F(1,124)=1.246, ns$]. As expected, ASR quality had a significant effect on edit time, with low quality transcripts requiring additional edit times nearly three times the length of the audio (Mean=84.1s, One sample t-test, $p<.0001$). Average task time for low quality transcripts was 190.5s (approximately 6 times the average message length) and for high quality transcripts, 106.4s (approximately 3 times the average message length) [$F(1,124)=53.94, p<0.00001$]. However, editor interface did not appear to interact with ASR quality in overall task time.

Upon further investigation we noticed that there was a large disparity in the typing speeds of the participants in the initial typing test; participants cleanly divided into two clusters. *Fast typists* had typing speeds ranging from 51 to 74 correct words per minute (cwpm), while *slow typists* had typing speeds of 25 to 35 cwpm. The fast typist group contained 10 participants (median 65.5 cwpm) and the slow typist group contained 6 participants (median 31.5 cwpm).

Looking at the two new groups, we found that slow typists were faster in the menu condition than in the swipe condition (menu=153.45s, swipe=160.35s) but fast typists performed better in the swipe condition, taking 133.75s compared with 155.48s for the menu condition. Neither effect was significant. Note that both groups' performance was nearly identical in the menu condition, but very different for the swipe condition.

In addition to task time, we also wanted to evaluate the amount of effort for each task. We defined a metric called *edit operations* to represent task effort in the two interfaces. Edit operations consisted of the number of word substitutions plus any additions or deletions. For example, “*less in this room*” changing to “*our fitness room*” includes three edit operations: two substitutions and one deletion. A summary of edit operations is shown in Table 3. In all conditions, participants performed more edit operations in the menu condition than the swipe condition, indicating that they accomplished more during the task. The differences in ASR quality [$F(1,124)=69.35, p<0.00001$] and editor [$F(1,124)=4.356, p<0.05$] are significant, but the differences across typing group were not [$F(1,124)=2.73, ns$].

		Edit time per message (s)		
		Menu	Swipe	Average
ASR Quality	High	92.2	69.9	81.0
	Low	179.5	130.7	155.1
	Average	135.9	100.3	

Table 2: Participants spent significantly more time using the menu editor than swipe-and-type. Note these values are after subtracting audio playback time.

		Number of edit operations per message (in words)		
		Menu	Swipe	Average
ASR Quality	High	7.8	7.4	7.6
	Low	23.0	17.1	20.1
	Average	15.4	12.3	

Table 3: Participants edited more words with the menu than with swipe-and-type.

However, this may not give an entirely clear picture of task effort. Since the messages we used in the experiment were real-world conversational speech, we needed a way to evaluate the importance of words in the transcript to the comprehension of the message. One possible method is to label each word in the transcript as either a function word or a content word. Linguists often draw a distinction between function words, those words with little lexical meaning but important roles in the grammar of a language, and content words, whose meaning is best described in a dictionary.

Before looking at function/content words we fixed user spelling errors, expanded contractions, and applied a Porter stemmer to all words in the corrected transcripts. The analysis shows that in the menu condition, participants in both groups edited far fewer function words than in the swipe-and-type condition (41.33 function words for menu versus 68.15 function words for swipe-and-type [$F(1,126)=8.575, p<0.01$]). In the menu condition, more of the edits were of important (content) words. If we look at all of the words edited across all tasks, 52% of the words slow typists edited in the menu condition were content words compared to 42% in the swipe condition. Fast typists had similar results, 46% content words edited in the menu condition and 43% content words in the swipe-and-type condition [Editor: $F(1,124)=4.19, p<0.05$; Typing group: $F(1,124)=3.9, p<0.05$].

An interface that requires typing has the potential to cause more spelling errors than a menu-based interface. We found 37 total spelling errors made by 14 of the 16

participants across all but one message. Two participants made no errors. Of the total spelling errors, 29 were made using the swipe-and-type editor, but only 8 were made using the menu (in typing a word not on the menu). Participants made significantly more spelling errors with swipe-and-type [$F(1,126)=9.21, p<0.01$].

Subjective Results

Participants were undecided about which editor was easier, generally saying it depended on whether errors were single words or longer phrases. On a 5-point Likert scale (5=easy), fast typists rated swipe-and-type as 3.8 and the menu as 3.0. Slow typists responded with similar ambivalence, rating swipe-and-type as 3.0 and the menu as 3.33. When asked whether the menu was faster than swipe-and-type, participants disagreed (2.25 on a 5-point Likert scale with 0=disagree, One sample t-test, $p<.05$). Participants also marginally disagreed with the statement that the menu was easier than swipe-and-type (2.63, One sample t-test, $p<0.1$).

Previous studies have suggested confidence shading does not improve time to locate errors [17, 6], but participants in the present study agreed with the statement “the gray-colored text was helpful for identifying mistakes in the transcripts” (4.06 with 5=completely agree, One sample t-test, $p<.001$).

User Comments

In general, participants liked SCANMail (“Cool software!” “How can I get this?”) and said they would like a combination of both swipe-and-type and menu editors. In the current study, the majority preferred swipe-and-type over the menu editor because errors usually extended to multiple words in a row: “Swipe-and-type is better for the (frequent) cases where whole phrases need to be changed, including word segmentation errors.” Editing individual words with the menu was tedious: “The menu didn’t allow me to get rid of words I didn’t want quickly enough.” However, the menu supported quick word changes without pausing the audio: “It allowed me to follow the message and make corrections quickly as it played.” In general, most participants expressed a general preference for the keyboard over the mouse: “I’m a keyboard person more than a mouse person.”

By attending more to content than function words within the menu editor, some participants disliked leaving incorrect function words scattered throughout the text: “It would be hard to scan visually later.” However, others suggested the menus helped them to focus on important words: “Otherwise I would just end up correcting people’s grammar.”

Participants noted one particular benefit of the menu, especially in low quality transcripts: Having a list of possible words was helpful in determining the correct word without playing audio: “I can tell what the computer thinks

and probably it will help me.” Audio playback times were, indeed, shorter in the menu condition.

DISCUSSION

The present study strongly enforces the idea that people want to handle voicemail similarly to email, but current voicemail systems hinder their ability to do so. SCANMail improves on voicemail by displaying text transcripts for quick visual skimming and long-term archive. Empirical evaluation of typing- and menu-based transcript editors shows that the menu leads to shorter audio playback times and an emphasis on editing *important* words, but increased message edit time. The study raises several issues related to the editors’ efficacy.

First, the significant reduction in audio playback time along with minimal keyboard interaction for the menu-based editor suggests that the menu may be valuable in mobile environments where limited input capabilities are the norm and user privacy is essential. Presenting a list of alternative words reduces the amount of conspicuous audio playback required. However, additional improvements to the menu interface will be necessary to reduce the time required per word edited. Restricting menu edits to content words may help to focus users on important errors and supporting phrase-level menu edits could potentially ease the complex interactions required for sequential word-level editing.

Second, the nature of the errors has a dramatic effect on editor use: Phrase-level edits are faster to correct by typing. Confirming previous findings that errors tend to cluster [6], multi-word errors were the norm, even in high quality transcripts. With swipe-and-type, participants could wrap more words—including unimportant function words—into a single swipe, rather than replacing them one at a time. Faster typists especially benefited from this method, in accordance with previous studies [17]. Though overall performance speeds did not differ dramatically across editors, the participant survey responses strongly indicate a preference for swipe-and-type because of the number of phrases needing repair.

However, the menu-based editor had specific advantages. Spelling errors were minimized, an important consideration for long-term archiving of messages. Also, slower typists did tend to perform better with the menu than with swipe-and-type. The typing speeds within this study—even those of “slow” typists—were faster than those in the general population; testing the menu editor with a more diverse demographic would reveal its usefulness for slower typists.

Both editors had limitations. Users found word-level editing tedious with the menu, and many resorted to selecting “other” and typing replacement text without skimming the menu first. With swipe-and-type, phrase-level replacement had the potential to cause alignment errors between the text and the audio, so users were limited to selecting eight words at a time. Sometimes the phrases in low quality transcripts were one or two words longer

than the selection constraint, requiring users to perform two separate replacement actions. Understanding the limitations of both editors will better inform the design of error correction on mobile devices. Furthermore, the full version of SCANMail mitigates many of these limitations artificially imposed within the lab environment. SCANMail produces high quality transcripts by default and supports both menu- and swipe-and-type editing within a single message.

To control message content and avoid privacy issues, we presented participants with novel voicemail messages from callers they did not know. Messages contained general content that did not require external context, but personal messages from friends and coworkers would most likely affect editing behaviors. Knowing how someone speaks may make message transcripts easier to understand, even transcripts with many errors. It is not known whether people would play less audio or make fewer corrections to messages from familiar people.

This experiment adds to the scarce body of empirical knowledge about the usefulness of confidence shading for locating errors. First, it suggests that the confidence threshold matters: Globally minimizing categorization errors increases the number of missed detections, which may lead users to miss critical misinterpreted words. Increasing the confidence threshold presents more gray words, serving as an implicit reminder that transcripts may have mistakes and should not be automatically trusted. Participants said the gray words were easy to visually skim, and that “*it became ingrained that they might be wrong.*” In the present study, we did not test whether confidence shading improved error detection *speed*, simply whether users found the shading helpful. Unlike commercial dictation systems, voicemail messages are brief and need only sporadic editing, so perceived usefulness may be more important than exact speed. The accuracy of the confidence shading was increased slightly (3.4%) in this study to maintain consistency across the low and high quality transcripts, suggesting that with minor improvements in ASR confidence scoring, confidence shading may indeed be useful.

CONCLUSION

SCANMail brings an email-like interface to voicemail, overcoming many of the limitations inherent in audio-only applications. Text transcripts serve as a visual analog to the speech data, allowing users to easily browse for relevant information and permanently archive messages. However, speech recognition errors limit the usefulness of the transcripts, especially for keyword searching and long-term comprehensibility. Two interfaces for user-initiated error correction were studied, revealing that the grammar-assisted menu editor led to lower audio listening times and an increased number of important words edited.

Future Directions

Participants suggested other editor interfaces, including phrase-length menus and progressive completion extensions to swipe-and-type. SCANMail now supports the replacement of three-word phrases with menus; the word confusion network suggests the best trigram paths through the lattice for each point in the transcript. Progressive completion would suggest terms based on the first few characters users type, and is popular in other applications for frequently occurring or complicated strings, such as URLs. Suggested terms would come from the WCN or a general dictionary.

SCANMail runs on numerous mobile devices. Most current SCANMail users check their voicemail transcripts at their office computers, so the desktop computer tested here was an appropriate platform. With a full-sized keyboard, users’ preference for swipe-and-type is not surprising. The success of the menu editor in reducing audio playback time indicates that it will be a viable interface for error-correction on mobile devices.

The present study contributes to the much-desired integration of communication technologies, such as cell phones, PDAs, and laptops, presenting information in a format appropriate to the platform and in a way that humans process most naturally. SCANMail allows email and voicemail to be handled similarly, transitioning voicemail from a cumbersome burden that is promptly deleted to a long-term informational resource that can be retrieved later on any device.

ACKNOWLEDGMENTS

We would like to thank Anthony Hornof, Bob Kraut, and Scott Hudson for their suggestions on previous drafts, and Maria Velazquez and Bob Bell for their assistance with experimental design and analysis. Thanks also to our voice actors and experiment participants.

REFERENCES

1. Arons, B. SpeechSkimmer: A system for interactively skimming recorded speech. *ACM Transactions on Computer-Human Interaction* 4, 1 (1997).
2. Bacchiani, M., Hirschberg, J., Rosenberg, A., Whittaker, S., Hindle, D., Isenhour, P., Jones, M., Stark, L. and Zamchick, G. SCANMail: Audio navigation in the voicemail domain. *Proc. Conference on Human Language Technology Research 2001*, ACM Press (2000), 1-3.
3. Boreczky, J., Gigensohn, A., Golovchinsky, G., and Uchihashi, S. An Interactive Comic Book Presentation for Exploring Video. *Proc. CHI 2000*, ACM Press (2000), 185-192.
4. Chase, L. Word and acoustic confidence annotation for large vocabulary speech recognition. *Proc. Eurospeech 1997*, (1997), 815-1818.

5. Degen, L., Mander, R., and Salomon, G. Working with Audio. *Proc. CHI 1992*, ACM Press (1992), 413-418.
6. Feng, J. and Sears, A. Using confidence scores to improve hands-free speech based navigation in continuous dictation systems. *ACM Transactions on Computer-Human Interaction*, 4,11 (2004), 329-256.
7. Hakkani-Tür, D., Béchet, F., Riccardi, G. and Tür, G. Beyond ASR 1-Best: Using word confusion networks in spoken language understanding. *Journal of Computer Speech and Language*, Elsevier, (To appear).
8. Hauptmann and Witbrock, M. Informedia: News-on-demand multimedia information acquisition and retrieval. *Intelligent Multimedia Information Retrieval*, AAAI Press (1997), 213-239.
9. Hazen, T., Polifroni, J., and Seneff, S. Recognition confidence scoring for use in speech understanding systems. *Computer Speech and Language* 16, (2002), 49-67.
10. Hindus, D., Schmandt, C., and Horner, C. Capturing, structuring, and representing ubiquitous audio. *ACM Transactions on Information Systems* 11, 4 (1993), 376-400.
11. Karat, C., Halverson, C., Karat J., and Horn, D. Patterns of entry and correction in large vocabulary continuous speech recognition systems. *Proc. CHI 1999*, ACM Press (1999), 568-575.
12. Kazman, R., Al-Halimi, R., Hunt, W., and Mantei, M. Four paradigms for indexing videoconferences. *IEEE Multimedia* 3, 1 (1996), 63-73.
13. 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. *Proc. CHI 1997*, ACM Press (1997), 202-209.
14. Oviatt, S. Taming Recognition Errors with a Multimodal Interface. *Communications of the ACM* 43, ACM Press (2000), 45-51.
15. Stark, L., Whittaker, S., and Hirschberg, J. ASR satisficing: The effects of ASR accuracy on speech retrieval. *Proc. International Conference on Spoken Language Processing*, (2000).
16. Stifelman, L., Arons, B., and Schmandt, C. The audio notebook: Paper and pen interaction with structured speech. *Proc. CHI 2001*, ACM Press (2001), 182-189.
17. Suhm, B., Myers, B. and Waibel, A. Multimodal error correction for speech user interfaces. *ACM Transactions on Computer-Human Interaction* 1, 8 (2001), 60-98.
18. Vemuri, S., DeCamp, P., Bender, W., and Schmandt, C. Improving speech playback using time-compression and speech recognition. *Proc. CHI 2004*, ACM Press (2004), 295-302.
19. Whittaker, S. and Amento, B. Semantic Speech Editing. *Proc. CHI 2004*, ACM Press (2004), 527-534.
20. Whittaker, S. and Amento, B. Seeing what you are hearing: Co-ordinating responses to trouble reports in network troubleshooting. *Proc. ECSCW*, Kluwer Academic Publishers (2003), 219-238.
21. Whittaker, S., Davis, R., Hirschberg, J., and Muller, U. Jotmail: A voicemail interface that enables you to see what was said. *Proc. CHI 2000*, ACM Press (2000), 89-96.
22. Whittaker, S., Hirschberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., Stead, L., Zamechick, G. and Rosenberg A. SCANMail: A voicemail interface that makes speech browsable, readable, and searchable. *Proc. CHI 2002*, ACM Press (2002), 275-282.
23. Whittaker, S., Hyland, P., and Wiley, M. Filochat: Handwritten notes provide access to recorded conversations. *Proc. CHI 1994*, ACM Press (1994), 271-277.
24. Wilcox, L., Chen, F., Kimber, D., and Balasubramanian, V. Segmentation of speech using speaker identification. *Proc. International Conference on Acoustics, Speech, and Signal Processing* (1994), 161-164.