iRemember: a Personal, Long-term Memory Prosthesis

Sunil Vemuri, Chris Schmandt, Walter Bender
MIT Media Lab
20 Ames St.
Cambridge, MA 02139 USA
{vemuri, geek, walter} @media.mit.edu

ABSTRACT
We present a wearable, computational memory aid capable of ubiquitous recording and associated retrieval tools for use during memory failures. We describe a study in which one of the authors recorded everyday conversations with colleagues for two years and subsequently evaluated the effectiveness of the retrieval tools for remedying simulated memory problems. Results suggest early validation of the memory retrieval approach (i.e., searching for memory triggers) towards alleviating certain classes of memory problems.

Categories and Subject Descriptors
H.5.2 User Interfaces, H.3.3 Information Search and Retrieval.

General Terms: Human Factors, Experimentation

Keywords
Memory aid, Speech Recognition. Information retrieval

1. INTRODUCTION
Imagine being able to remember better; remembering names, facts, ideas, conversations, etc.; being able to vividly reminisce about fond memories. Unfortunately, human memory can be ephemeral, fallible, and malleable; it falters at the most inconvenient times and circumstances. Failing memory can be devastating to a person's productivity and psyche and memory failure, bias, and especially manipulation can have serious legal and public safety repercussions [5],[9].

Ubiquitous computing (e.g., via portable computing devices) offers compelling opportunities to mitigate some memory shortcomings by supplementing one's biological memory with the verbatim, unbiased, and unfiltered recordings of life events stored in computer memory. The notion of "mechanical" memory assistance is not new; it was proposed nearly sixty years ago in Vannevar Bush's "Memex" article [2]; over the subsequent years, various efforts have been made to realize this vision.

In this paper, we present iRemember: a computer-based memory aid capable of assisting with some everyday memory problems. It does this by ubiquitous recording (primarily audio) of everyday conversations, transcribing these with automatic speech recognition (ASR), and making recordings available as a browseable and searchable resource; users can turn to iRemember anytime they need help remembering. To better understand how such an approach can help remedy some everyday memory problems, we conducted a two-year study in which one of the authors recorded everyday conversations with colleagues who evaluated how effective iRemember was in helping resolve simulated memory problems.

We call this approach memory retrieval (MR): information retrieval (IR) is used to find memory triggers. While the underlying technologies of MR and IR may be similar, the success criteria differ. In IR, a successful result is one that contains the sought-after information. In MR, a successful result is one that either contains the sought-after information or triggers the memory of the information. It is easier to build such a system than to evaluate since the task, by definition, requires a long-term perspective of at least several years of data.

In addition to being a position paper, our previous paper [15] gave an overview of the evolution of the same system including synopses of the previous and present experimental results. This paper provides additional significant results of our two-year study and a more-detailed description of its evaluation methodology. The remainder of the paper provides background on computational memory aids, describes the iRemember memory prosthesis, and then details the evaluation we performed to better understand the efficacy of the MR approach and its suitability to certain memory problems.

2. BACKGROUND
Memory aids can take the form of strings on fingers, sticky notes, mnemonics, etc. The present focus is on computational memory aids. Before describing technological approaches, we start with a brief discussion of memory problems.

Schacter's taxonomy, the "Seven Deadly Sins of Memory," succinctly describes the most-common memory problems [12]. The six involving forgetting and distortion are shown in Table 1. The seventh, "persistence" (pathological inability to forget), is of less interest to memory-aid designers. Studies of workplace memory problems indicate that transience and absent-mindedness are the most oft-cited problems [3].

Table 1: Six of the seven "sins of memory"

Forgetting	Distortion
Transience (memory fading over time)	Misattribution (right memory, wrong source)
Absent-mindedness (shallow processing, forgetting to do things)	Suggestibility (implanting memories, leading questions)
Blocking (memories temporarily unavailable)	Bias (distortions and unconscious influences)

CARPE'06, October 28, 2006, Santa Barbara, California, USA.

iRemember aims to address transience. The approach is to collect, index, and organize data recorded from a variety of sources related to everyday activity, and to provide a computer-based tool to both search and browse the collection. The hope is that some fragment of recorded data can act as a trigger for a forgotten memory.

It is anticipated that blocking problems would also benefit from such an aid. One of the common qualities of both transience and blocking is that the person is aware of the memory problem when it occurs (this is not true for all memory problems). Assuming the person also wishes to remedy the problem, what is needed is a resource to help. This is where iRemember comes into play.

Regarding long-term, personal memories, Wagenaar [18] and Linton [8] performed separate, influential multi-year diary studies in which they recorded salient experiences every day in a written journal. For six years, Wagenaar wrote down the time, place, who was with him, and a brief statement about the daily events. At the end of the recording period, an assistant tested his recollection of randomly selected episodes. Linton performed a similar study. The significance of these works rests in their magnitude and application to real-world memories (with the caveat that retrieval was performed in the laboratory). Wagenaar's experiments illustrate a sharp decay in recall over the first year and then a steady decay afterward. Furthermore, his results suggest more retrieval cues leads to better retention. Finally, both Wagenaar and Linton illustrate useful methodologies for longitudinal self-evaluation of memory.

2.1 Memory Aids and Personal Data Archival

Sixty years ago, Vannevar Bush postulated storing—in a giant associative memory—the documents used over a lifetime along with the trails and histories of work done in the process [2]. This archive, combined with a "mechanized" retrieval system, was a proposed memory aid called Memex. Years later, technological advances enabled serious attempts to realize this vision. Some manifestations included tools to capture or record activities in rooms with cameras and electronic whiteboards [1],[10]. Toward the end of the 1990s, approaches using portable devices [6],[11],[13] started to become feasible. The ubiquitous recording device is attractive because life is not limited to a conference room and many significant daily activities—or at least activities we may wish to recall later—do not occur on a scheduled basis or at a particular location.

Now, the ability to keep verbatim records of one's life experiences is technologically possible and affordable. Research and industry efforts are actively working on personal-archival tools [4],[7]. Archival features in personal-computer communications applications (e.g., text messaging, voice chat, etc.) are becoming increasingly commonplace. Portable devices (e.g., PDAs, smartphones, etc.) are rapidly decreasing in cost, and consequently, are now valuable deployment vehicles for ubiquitous recording projects. Storage is inexpensive; high-speed wireless networking is ubiquitous.

The designers of archival systems (including us) are keenly aware that these tools, while well-designed and architected for capture and storage, may result in "write once, read never" repositories. Some reasons include the limited- and distant-benefits motivating archival. Also, the rapidly improving technology for capture has surpassed the tools to assist with organization and retrieval of collected data. Regarding the latter, key enabling technologies such as automatic, large-vocabulary speech recognition and IR engines now run on commodity personal computers. Yet, ASR still suffers from high error rates and IR focuses mainly on text data, not on image, audio, or video retrieval. Furthermore, these techniques tend to treat data from personal experiences no differently than "generic" data.

To address the former, we associate the purpose of everyday archival with solving everyday problems. Specifically, we are interested in how such archives can be used for everyday memory assistance. Our approach has centered on audio due to the ease of capture, availability of inexpensive, portable recording equipment, availability of suitable indexing technology, tractable storage requirements, and value of audio as a memory trigger. Once collected, we employ searching and browsing tools to help users find these triggers.

3. The iRemember "Memory Prosthesis"

The basic architecture for iRemember includes a recording apparatus and associated retrieval tools. To facilitate capture of life experiences in a variety of everyday situations, we developed software for use on the iPaq 3650 PDA[1] (Figure 1). The device is capable of "speech-recognition-quality" audio recording via a built-in near-field microphone. We modified the device to also accept an external lavalier microphone to increase convenience, aesthetics, and improve audio quality. Low-cost contemporary solutions are available that afford even higher-quality microphoning. For example, headset noise-canceling microphones can virtually eliminate secondary speakers [20]. We opted for the built-in iPaq microphone and a lavalier microphone since we wanted audio from secondary speakers; preliminary, informal studies suggested secondary speaker inhibition when conversing with someone wearing a headset microphone.

Due to storage limitations on the PDA, the software records audio and immediately transmits these via a high-speed wireless network to a large-capacity server where all audio recordings and associated data are archived. Speech recognition is also done on the server using IBM's ViaVoice [17].

Once data are collected on the server, users can browse and search using software we developed for both the PDA and conventional, personal computers. The present evaluation is based on use of the personal-computer-based application. Details of the PDA-based retrieval tools can be found elsewhere [14]. The personal-computer retrieval tool (Figure 2) allows for browsing and searching of the entire collection. Additionally, it shows the recordings in the context of other data at the time of the recording. This includes the location of the recording, calendar events, email, and local weather at the time. Users can conduct keyword searches to find information in any of the recorded data. Results are shown as a ranked list and on a timeline (Figure 3).

[1] The iPaq 3650 was one of the only devices capable of the technical requirements at the project's onset. Many contemporary devices are more capable and would be better suited for subsequent evaluations.

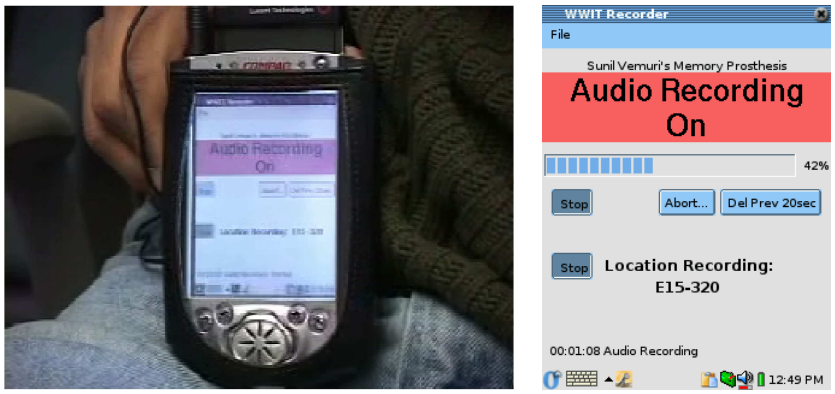

Figure 1: iRemember capture software running on iPaq PDA (top). Closeup of screen on the bottom

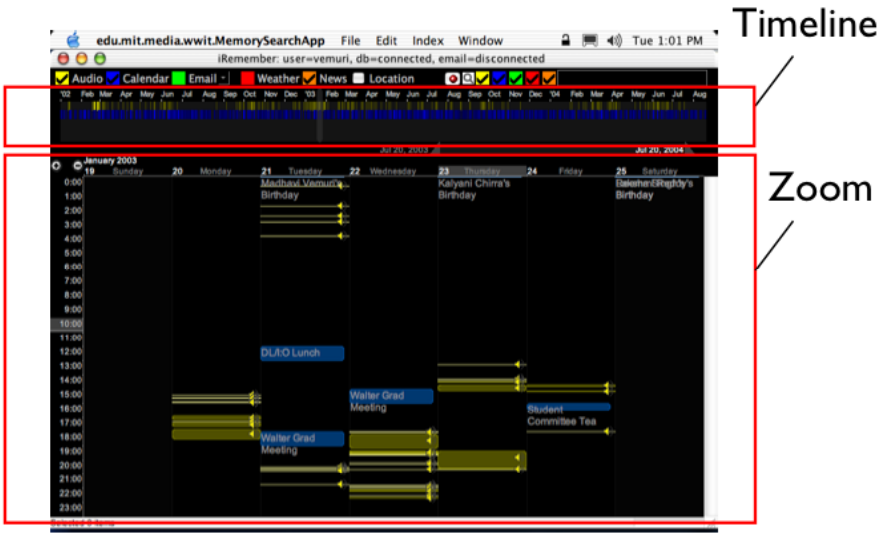

Figure 2: Visual interface for browsing and searching through all recordings. This view shows a multi-year timeline on the top and a zoomed-in view of one week on the bottom

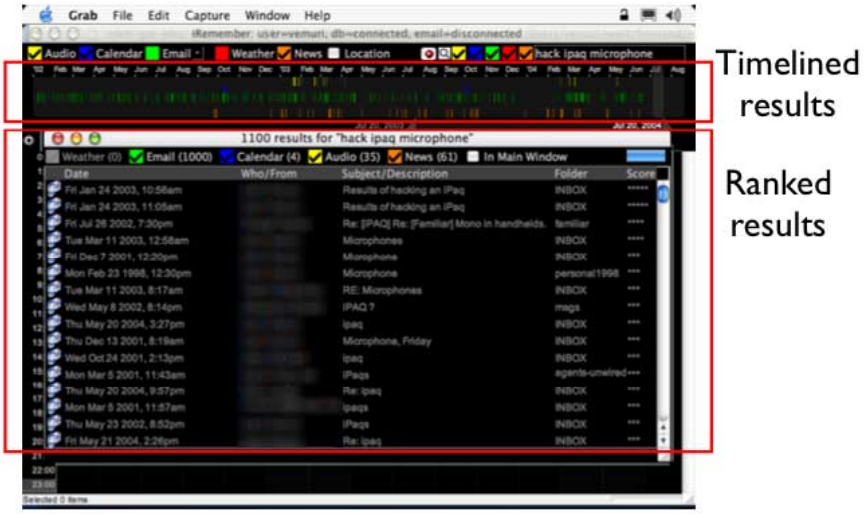

Figure 3: Keyword search results shown as a ranked list and on the timeline

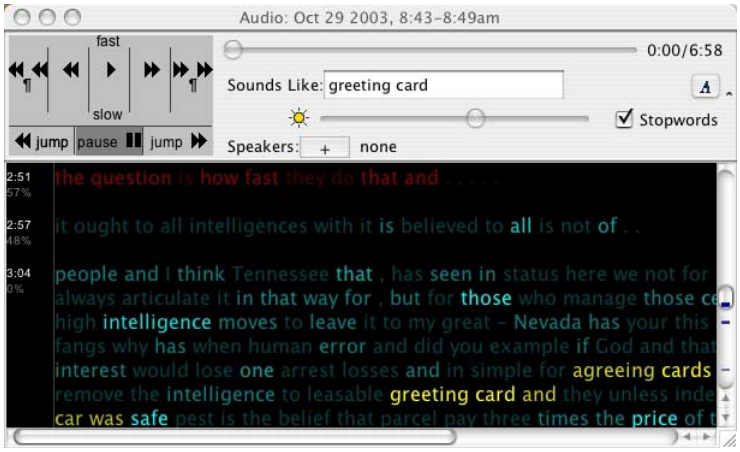

Figure 4: Interface to browse and search an individual recording

Users can double-click on an audio recording from either UI to open another viewer (Figure 4). This shows additional detail and provides the ability to play the audio, see the speech-recognizer-generated transcript, search within the data, etc. Transcript text brightness is proportional to recognition confidence. A phonetic or "sounds like" search feature was included to reduce the effects of transcription errors.

4. EVALUATION SETUP AND METHODOLOGY

In brief, one of the authors (RA for "recording author") recorded conversations with departmental colleagues over a two-year period. Not every conversation was recorded for a variety of reasons (e.g., privacy, investigator choice, absent-mindedness, etc.). After the recording phase, the RA listened to all of the recordings, constructed questions, and asked three recordees to recall information contained in the conversations. When the recordees could not remember the answer, they could use iRemember to browse and search through the data to find the answers while the RA observed their retrieval efforts.

Challenges and confounds in long-term evaluations of such a tool are inevitable. Low-accuracy speech recognition is a given for any real-world deployment; issues such as microphone placement, environmental noise, interpersonal distances, aesthetics of an always-worn recording apparatus constrain the best speech regonizers available. We wanted the evaluation to be as realistic as possible. Subject recruitment constraints, selection, and bias also contribute imperfections but may be inevitable for an in-depth, multi-year study with budgetary constraints. Despite this, we felt it was instructive to better understand the nature of these challenges and evaluate the validity of the approach. This remainder of this section details conversational-data collection procedures, subject selection, question construction, and the memory-retrieval test.

4.1 Conversational Data Collection and Subject Selection

Collecting data from personal conversations poses challenges. First, unlike lecture and meeting situations—which are increasingly being recorded for archival and related purposes—day-to-day conversations are typically not recorded. Second, although availability of suitable portable recording devices is commonplace, social conventions still hinder such recording behavior in most non-experimental situations.

Another challenge of long-term evaluations is subject selection. The commitment is long and it is difficult to secure a representative subject pool. Self-tests—like Wagenaar's and Linton's—have advantages and were considered. The RA could be tested in a similar manner (e.g., ask an assistant to administer questions). Confounds related to weaknesses and subtleties in the tools (especially usability issues) could be minimized. Unfortunately, doing so would require that the RA not use the multi-year collection of recordings for day-to-day memory needs (which he was doing) and would result in N=1. So instead, he tested others: none of who had the retrieval software or access to the recordings during the recording phase.

Three colleagues (Subjects A, B, and C) were chosen to participate in memory-retrieval tests; these were selected primarily due to the number of conversations the RA had with them over the two-year data-collection period. Other colleagues were recorded in this span, but only a few were recorded with sufficient frequency and duration to provide enough data for this study. Subjects had normal occasion to converse with the RA on a regular basis (i.e., outside of this study), were interested in the research, and were sympathetic to the work. Only colleagues who felt comfortable with audio recording and archiving volunteered. All subjects were advanced computer users (10+ years of experience) with prior exposure to speech recognition, information retrieval, and evaluation techniques.

4.2 The Personal Experiences Data Set

The audio recordings and location data were collected using the iRemember wearable recording apparatus (Figure 1). Table 2 shows some basic statistics about the audio recordings between each of the subjects and the RA. In a few cases, two of the subjects and the RA were in the same conversation; these are counted in the tallies for both subject. No recordings included all three subjects and the RA.

As expected, the speech-recognizer-generated transcripts for the data set suffered from high word error rate[2] (WER). Some

[2] WER = (insertion + deletion + substitution errors)/number of words in the perfect transcript

sections were better than others; variability depended primarily on the proximity of the speaker to the microphone. With only one microphone, placement was a challenge. Speech recognition performs better with a high-quality signal from a near-field noise-canceling microphone. Various microphones and configurations were tried. After factoring in aesthetics and convenience, the preferred option was positioning a near-field non-noise-canceling microphone (i.e., iPaq built-in or lavalier) close to the primary speaker (~70% WER for the primary speaker and ~100% for other speakers) instead of midway between speakers (95%+ WER for all speakers). Most errors for secondary speakers were deletion errors; hence, false positives were not common. For most conversations, the RA was the speaker closest to the microphone, but he occasionally placed the device or external microphone either close to the subject or midway between the speakers. WER was generally highest when either the distance between the speaker and the microphone was more than a few feet or multiple people spoke simultaneously. The estimated WER for transcripts generated from the RA's uninterrupted speech is 70%. Interrupted speech resulted in a higher WER.

Table 2: Basic statistics on recordings between each subject and recording author (RA)

	Number of recordings	Mean duration (minutes)	Median duration (minutes)	Total time (hours)
Subject A	58	10.2	6.2	9.8
Subject B	58	9.3	4.0	8.9
Subject C	45	11.7	7.6	8.7

The estimated WER for the secondary speaker was worse (~100%). Most of those errors were deletions that could be attributed to the low recording amplitude commensurate with the distance between the secondary speaker and the near-field microphone: so few words from the secondary speaker mean fewer false positives but more false negatives. With better microphone conditions, one would expect the WER to resemble the primary speaker's. Although the speech recognizer had high WER for secondary speakers, the speech was still somewhat audible to a human listener. For memory retrieval, some audio is better than none. Future studies might benefit from all conversation participants wearing near-field noise-canceling microphones. For the present study, an adequate microphone setup was the goal.

Only one speech-recognizer voice model (the RA's) was used for all speakers. Under these circumstances, one would expect WER for the RA's speech to be better than other speakers. However, this was not found to be true.

Aside from the audio recordings of conversations, subjects were asked to provide their entire calendar and email archives covering the entire two-year recording period. All subjects maintained email archives over the entire span; Subjects A and B maintained calendar data during this span; Subject C archived data only for the last year but said he would archive calendar data longer if he became a regular user of the tool.

News reports were archived automatically every night by capturing the main page of several popular news websites (CNN, New York Times, Google News, etc.); these were provided to each subject during the experiment. Weather data, including textual descriptions for the local area were collected on an hourly basis throughout the two-year data collection period and these were also provided.

Sometimes there were few recordings in a wide time span (e.g., one recording in a month). The RA may have only spoken with that colleague once that month, only recorded one conversation, or did not record the conversation for other reasons. In the task, if subjects could narrow their search to that month, the task of localizing within the collection was simplified. The nearly always-on ubiquitous recording vision suggests far more data would be recorded; presumably, the higher the recording density, the harder the task. Conversely, assuming the RA's behavior is typical, the volume is a reasonable approximation.

4.3 Question Construction

Questionnaire construction is a time-consuming process. It takes roughly four times the duration of a recording to listen, re-listen, and extract some meaningful questions for a memory test. The RA listened to every conversation, identified some potentially interesting passages, and phrased them into questions. Questions were on topics typical of the everyday conversations between the parties and relevant to research tasks and personal interest.

Questions were designed to try to evoke transience memory problems; that is, questions that the subjects probably knew the answer at some point in the past, probably would not know the answer during the test due to memory fade, and would need assistance to answer it. In this sense, they were all "hard" questions. There was no way to know if the subject would experience a memory problem until the question was given. Also, there was no way to know if the subject actually encoded the relevant information to long-term memory at some point in the past. To estimate this, question topics were selected based on careful listening to the original recordings to see if there was some indication that the memory achieved some form of long-term encoding. Examples of such indications include: the subject originally spoke the answer, the subject engaged in a conversation related to the answer, the subject asked a question related to the answer, or expressed some other indication of interest in the information in the original conversation.

Questions were designed so subjects would give free-form answers. There were no true/false or multiple-choice questions. All questions had an answer that could be found among the recordings and questions were designed such that the answer was unambiguous and the entire answer could be found within a short span (roughly 30 seconds) of a single recording. In this sense, the questions were biased towards subjects' episodic memories versus their semantic memories. Subjects could ask for clarifications of the question. This happened on several occasions. A few sample questions are shown in Figure 5.

Questions are biased towards statements made by the RA due to the better recording quality, and hence greater speech-recognition accuracy. After phrasing the question, the RA used various combinations of words from the question to see if keyword and phonetic searching alone could be used to retrieve the answer. By design, most questions could (~90%); some could not.

- What computer techniques were used to help sequence the genome?
- Which Star Trek episode does [RA] think is one of the best, which one is one of the worst?
- How does Singapore maintain ethnic balance in its population?
- What does MYCIN do?
- Who started the open source movement? What was the name of the project this person started?
- What did [RA] say one should do prior to visiting the Getty Museum?
- What is [RA's] opinion on the use of common-sense reasoning to help interpret speech-recognizer-generated transcripts? On what does he base this position?

Figure 5: Sample questions given to subjects

4.4 Testing Procedure

Subjects were presented with one "task question" at a time. Immediately after reading a task question, they were interviewed about their remembrance of the conversation, its context, and how they would approach the task of trying to remedy the memory problem assuming it was important to do so. The specific questions are shown in Figure 6.

For questions 2, 4, 5, 6, and 7, when a subject gave an answer, they were asked to assess their confidence in their answer on a scale of 0–10 with 10 meaning absolutely certain. If a subject answered Question 1 as "yes," answered the question correctly, rated their confidence in their answer high (8 or greater), and answered Question 3 as "yes," the subject was told their answer was correct and not asked to use the computer software to find the actual conversation. Under these conditions, the subject has demonstrated that there is no memory problem. Even if subjects could not answer Questions 4, 5, and 6 correctly, or gave incorrect information to Question 7, the question would still be classified as no memory problem.

After this interview, subjects were asked to use the memory-retrieval software to find the answer within the collection of recordings while speaking their thoughts out loud. The memory-retrieval software automatically logged all user interactions and the entire session was videotaped. Subjects had no time limit, but were allowed to give up at any time. Once the subject felt that they had found the answer or they wished to give up, they were interviewed again with a series of follow-up questions (Figure 7) similar to the earlier interview questions. At any time, subjects were allowed to give feedback and these comments were recorded.

The questioning took place over the course of two weeks with each subject sitting for multiple sessions. Subjects controlled the length of each session, which lasted from 10 minutes to two hours. In total, each subject spent roughly 4–5 hours attempting anywhere between 18–20 questions. Answering questions in the experimental setting can be both hilarious and fatiguing; when subjects were unsuccessful finding answers, they found it frustrating. One subject described the task as analogous to having someone create a Trivial Pursuit®-like game, but the category is always about you.

1. Do you remember having this conversation?
2. What would be your answer? (Guess if you wish)
3. Would you be satisfied with this answer if it were important?
4. When did this conversation take place?
5. Where did this conversation take place?
6. Aside from me [RA], who else was present?
7. Is there anything else you remember about this conversation?
8. Briefly describe how you would normally go about trying to find the answer, assuming it was important, but without the software I have provided.
   - What do you think are your chances of success this way?
   - How quickly do you think you would find an answer?
9. Now, assume [RA was not available *or* you could ask RA]. (The phrasing of this question depended upon what the subject answered for Question 8).
   - What do you think are your chances of success this way?
   - How quickly do you think you would find an answer?
10. What do you think your chances of success are using the software? How quickly do you think you can find the answer?

Figure 6: Questionnaire given to subjects after the main question, but before using the software

1. Do you remember having this conversation?
2. What is your answer?
3. Would you be satisfied with this answer if it were important?
4. When did this conversation take place?
5. Where did this conversation take place?
6. Aside from me [RA], who else was present?
7. Aside from what you just heard or saw, is there anything else you remember about this conversation?

Figure 7: Interview questions asked immediately after subject completed question-answering task

The RA assessed the correctness of the subjects' answers to the task question. An attempt was labeled unsuccessful if either the subject did not submit an answer or submitted an incorrect guess. An answer would be classified as correct even if the subject did not correctly answer any of the contextual pre- or post-questions.

The RA observed subjects throughout the attempt, identified memory problems, and classified these based on observations and subject verbalizations during the interview. The possible memory categorizations included all of Schacter's seven sins. Multiple memory problems were possible.

There were some unavoidable confounds. First, the process of answering one question could unintentionally improve a subject's memory and taint future question-answering attempts. As part of the normal question-answering process, subjects listened to verbatim audio from past conversations and reflected on their past. Both of these activities can strengthen memories of past events and surrounding circumstances, including events outside the bounds of the posed question. Such reflection can have the unfortunate effect of improving subjects' memories of past events as they progressed through the questionnaire. This would be reflected as improved question-answering accuracy and time-to-solution for later questions in the experiment. For a given conversation or segment of the conversation, the memory test can only be done once. Second, subjects were expected to become more facile with the memory-retrieval tool, the nature of spoken document retrieval, and the data set as they progressed. Performance on later questions was expected to benefit from this.

4.5 Retrieval Tools

Two slight variations of the personal-computer-based memory-retrieval tools were used as part of the evaluation, which will be called UI1 and UI2. The original intent was to use only one tool throughout the study. Part way through the study, subjects provided valuable feedback regarding features and improvements to the tool and user interface that they felt would improve their performance. Specifically, the first interface (UI1) did not allow phonetic searching across the entire collection of recordings; exact-match keyword searching was available across the collection, but phonetic searching only worked within an *individual* recording. Also, UI1 did not allow searches to be restricted to a limited date range. Usability tests on the lecture data trials [16] did not reveal these issues; studies on general spoken-document retrieval did not give much insight regarding personal data [19]. The importance of these issues were only uncovered in the present evaluation. This may be due to peculiarities of the personal-data situation or the higher WER among the recordings. These changes were incorporated into a second interface, UI2 (Figure 8), which was used in the remainder of the trials.

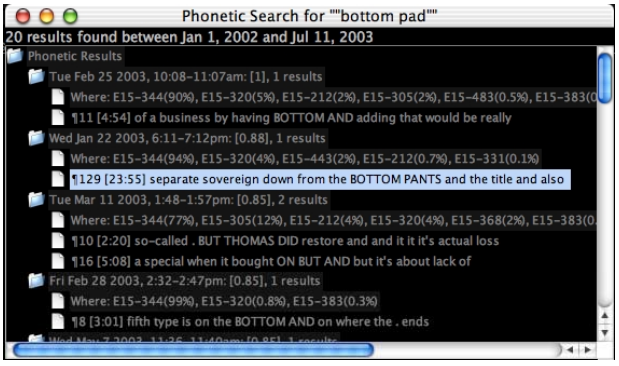

Figure 8: Exemplar results for collection-wide phonetic search

For the purpose of this task, subjects were not limited to using just the provided memory-retrieval tool. With the exception of asking the RA or anyone else, subjects were also allowed to use any other avenue to try to remedy the memory problem. For example, subjects could look for documents on their own computer, use their normal email or calendar client, search the web, items in their office, etc. Since the present focus is on memory remedies in the workplace via a computer, it was reasonable to expect that the computer might provide means outside of the memory-retrieval tool to identify landmarks.

5. Results

The results are broken into several categories: (1) memory and forgetting, but not to the software; (2) memory-retrieval in general, independent of the tool; and (3) how iRemember and its interfaces impacted a subjects' ability to answer questions.

In total, subjects attempted 56 questions. In 7 cases, subjects already knew the answer and did not have a memory problem. In 27 cases, subjects remembered having the conversation, but not the answer to the "task question." In 22 cases, subjects did not remember having the conversation, let alone the answer. This does not necessarily mean the memory of that conversation was not lurking somewhere in the subject's long-term memory, it just means that the question did not trigger a remembrance of the conversation. Table 3 summarizes the question-answering success depending upon whether the subject remembered having the conversation or not.

Table 3: Question-answering success depending upon whether the subject remembered having the conversation

	Remembered	Not remembered
Success (no software)	7 (13%)	-
Success (using software)	16 (29%)	13 (23%)
No success (using software)	11 (20%)	9 (16%)

A "successful" memory retrieval was one that resulted in a correct answer. An "unsuccessful" retrieval was one in which the subject either gave up during the attempt or stopped their search and provided an incorrect guess. It should be noted than all "incorrect-guess" cases, subjects essentially gave up on the task question and submitted a low-confidence guess.

Among the 49 task questions in which subjects had memory problems, transience memory problems were dominant with 45 cases; the remaining four were classified as misattribution problems. In 36 (64%) cases, subjects were successful in remembering either on their own (7 cases, 13%) or with help from the software (29 cases, 52%). Subjects succeeded even when they did not remember having the conversation (13 cases, 23%) and it was not surprising to see subjects did not succeed when they did not remember having the conversation (9 cases, 16%). Yet, it was disconcerting to observe the number of cases in which subjects remembered having the conversation, but could not find the answer (11 cases, 20%).

Figure 9 shows question-answering results over time. The x-axis corresponds to the amount of time that has passed between the original conversation and the memory test. The data are partitioned into two rows; the top row shows attempts in which the subject remembered having the conversation and the bottom row shows attempts in which the subjects did not remember having the conversation. Five of the seven cases in which subjects remembered the answer without using the software occurred within the six months prior to the test. Moreover, most of the no-success cases correspond to conversations that were over one-year past. This includes nine of the 11 cases in which subjects

remembered having the conversation, but were not able to find the answer.

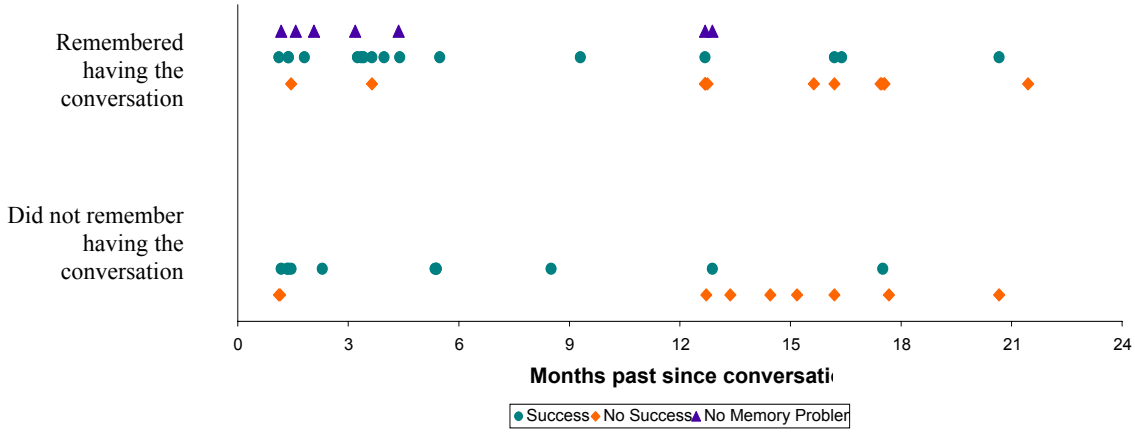

Figure 9: Question-answering attempts over time

Table 4: Subjects' prediction of time bounds. The number of questions are listed in each cell.

		Width of time prediction (months)								
		<1	**1**	**2**	**3**	**4**	**5**	**6**		**12+**
Correct prediction	Success	5	3	4	1		1	2		1
	No success		1		1			3		3
Incorrect prediction	Success		1		2					1
	No success		1		1			1		2

Better performance on more-recent memories is not surprising and it suggests that subjects' memories are aiding the retrieval process. To examine this in more detail, we look at how effectively subjects were able to predict the time bounds of a memory. For example, in the pre-questionnaire, subjects were asked when they thought the conversation took place. Subjects typically specified time ranges as a specific month, a range of months, a season of the year, a semester, etc. These were translated into a time width. If a subject said either Spring 2003 or Spring 2004, that would be a width of three months. If the conversation took place within the subject-specified bounds, that would be labeled as a "correct prediction;" otherwise incorrect. These were further partitioned into whether the subject succeeded in finding the answer to the task question. Tallies for this are shown in Table 4. The clustering of correct predictions corresponding to successes when the time width is two months or less gives further evidence that subjects' memories of the past events are helping with the memory-retrieval process.

While time-width predictions seemed helpful, confidence in answers was not so. Among the 49 questions in which subjects demonstrated memory problems, subjects submitted an incorrect guess and an associated confidence for 28 of these. There was no correlation found between a subject's confidence in their guess and the time passed since the event (r=0.22).

Based on the pre- and post-interviews, there was no evidence across subjects suggesting different remembrance of having the conversations. However, when Subjects A and C remembered having the conversation, they could cite specifics of the past conversations and the surrounding circumstances (e.g., who, where, what else). In contrast, Subject B's descriptions were more general and included references to multiple conversations with the RA on the asked-about topics or conversations on the topic with people outside of the study. This might suggest that Subject B's memories of these topics have become more consolidated as semantic memories whereas Subjects A and C are retained in episodic memory. The blurring of episodic details is not uncommon as memories become consolidated in semantic memory. This may explain Subject B's lower success rate and illustrate a limitation of the memory-retrieval approach for these types of memories.

5.1 User Interface Differences

As mentioned earlier, the change from UI1 to UI2 was at the behest of subjects who requested certain features in anticipation of better performance. In practice, subjects preferred UI2 and there were slight indications of increased question-answering success with UI2. However, there were too few examples (especially with the confounds) to make a strong claim that UI2 is better than UI1.

Table 5: Time spent per question with UI1 and UI2

	UI1			**UI2**		
	Mean	**Median**	**N**	**Mean**	**Median**	**N**
Time (success)	5:51	5:10	12	5:34	4:20	17
Time (no success)	10:38	8:07	12	13:14	13:45	8

A difference can be seen with the time spent per question for each interface (Table 5). Incorrectly answered questions and give-ups were lumped together since subjects were guessing and expressed low confidence in their answer. These results suggest that the second, interface (UI2) provides a small time-to-solution benefit. However, when subjects could not find the answer, they spent more time when using UI2. This might be because the subjects felt the newer user interface could do a better job with memory retrieval and were willing to give more effort to the task before giving up.

5.2 Qualitative Results

Subjects generally formulated an initial search strategy (e.g., keywords within a recording, looking for a landmark in either email or calendar entries, etc.), and tried it for some time. Initial failure typically led to minor variations on the query: for example, choosing slightly different keywords, variations on the original keywords, or different Boolean operators. If that did not work, they might try another path or give up.

Subjects primarily employed audio search as their first choice to remedy the memory problem. In fact, despite the ability to search email, calendar, news, and weather data along with audio, all subjects turned these features off at the beginning of each session. When asked why, subjects cited the length of time to get search results from all data sources, the expectation that the result would be found only in the audio, and the difficulty in navigating a large list of results. Except for a few isolated circumstances, subjects preferred to conduct email and calendar searches using their native applications. Below is a list of some other general observations from the present study related to search strategies:

- Not surprisingly, within-recording localization strategies were similar to those in our previous conference-talk memory-retrieval study [16] since the audio "document" visualization interface (Figure 4) was essentially unchanged between studies.
- Keyword search was the preferred mechanism for collection-wide audio search. This was contrary to the conference-talk study where calendar-navigation was the primary choice. With no temporal or landmark cue to use, keyword search is often the only remaining choice.
- Subjects employed simple mechanisms like calendar, email, and web search to find temporal landmarks.
- Accurate speech recognition makes the task easier.
- Less-vivid remembrance made it harder. This was based on subjects' answers in the pre-questionnaire indicating vivid details of the conversation, even if they did not remember the answer to the specific question.
- Misattribution when answering the "task" question, the pre-, or post-questionnaires led subjects astray (e.g., looking in the wrong time period).
- Subjects stated that multi-tasking was a helpful way to search through the data. For example, a subject would let an audio recording play while simultaneously reading another transcript or initiating another search.

5.3 Anecdotes

Several specific cases were instructive and entertaining. These are listed below in no particular order.

- When keyword searching of the main topic fails, subjects searched for what they thought was another topic in the same conversation. One subject remembered having a conversation, but was not successful in localizing within the collection using keyword search. Instead of giving up, the subject remembered another topic in that same conversation and started to search for keywords related to this second topic. The subject found the audio associated with this secondary topic and then skimmed the audio to locate the answer to the original task question.
- Three times, subjects remembered a conversation took place soon after a past seminar. The subjects searched their email (using both iRemember and their own email clients) for the talk announcement. Once found, they used that as a landmark and focused their attention on only the audio recordings soon after the talk.
- In one case, a subject remembered a conversation took place soon after the RA had returned from a conference. The subject used a web search engine to find the dates of a conference and focused on recordings occurring soon thereafter. Similarly, one subject remembered a conversation took place soon after his vacation and used that as a landmark.
- Some subject frustration in the study could be attributed to shortcomings in the user interfaces. The interfaces are research prototypes and admittedly had some usability shortcomings (e.g., speed, design, etc.). However, one subject also express frustration because of inability to produce the answer unaided "I should know this and I'm disappointed that I do not."
- In the most amusing question among them all, one of the subjects, expressing skepticism about the memory-retrieval approach, had stated two-years earlier that he felt it was "highly unlikely" that we would be able find this conversation years later; some colorful phrasing was used. Feeling up to the challenge, the RA phrased this into a question in that subject's test: "In the conversation where [Subject] and [RA] were talking about "[colorful phrase]," what did [Subject] say it was unlikely [RA] could do?" The subject found the answer.

6. Discussion and Conclusion

iRemember is an early step meant to demonstrate the efficacy of the memory-retrieval approach. The results illustrate that transience memory problems can be remedied via memory retrieval, despite high-WER transcripts. Subjects were able to achieve successes and when they did, answers could be found within a few minutes. There were times when the system could not help. One initially skeptical subject was amazed at how quickly he found obscure facts and stated, "When it works, it works great."

There are some caveats of the study. First, the memory problems were artificially created and this study does not shed light on whether subjects would want to commit to vigilant recording in order to reap the benefits. Some may consider a few minutes high; some may consider it low. In a time when finding information on the web takes seconds, a few minutes may seem long. Yet, if it is

important to remember, a few minutes can be incredibly short. The reader is invited to contemplate the types of daily memory problems you experience worthy of this effort. Second, there were not enough conversations. In some instances, there might have been only one recorded conversation in a given month and that could ease localization efforts. Assuming users are recording nearly continuously, the density of recordings would be higher.

Subjects' reactions were generally positive. Both during and after the testing, they became deeply engaged and expressed various reflections on the experience. These included personal introspection on their memory as well as comments on the technology. The subjects were interested enough to request new features (i.e., UI2), committed a large amount of time (4–5 hours testing), and made strong efforts despite the occasional frustration, and the fact that the memory problems were simulated.

The study will hopefully give insight to fellow researchers conducting similar evaluations. Performing such studies is time-consuming and it is important to get the details right at the onset. Once subjects are given a memory test and exposed to the data, their memory is refreshed and the data cannot be used again. Multi-year studies are still needed to gain insight into memory and we expect subject participation to remain low for these. During the data capture period, other volunteers expressed willingness to participate for shorter durations (e.g., a few weeks or months); they were recorded but the data were insufficient for the longitudinal study. In retrospect, testing these subjects would have been valuable towards usability studies and we may have received the UI2 and other valuable suggestions this way.

This study also explored the extent to which one's biological memory can aid in the memory-retrieval process. The success rates and time-bounding data suggest that one's memory helps mostly in the first year and diminishes afterward. This coincides with Wagenaar's results on decreasing recall during the first year [18]. Biological memory can also hurt: subjects occasionally experienced the "misattribution" memory problem during the initial search; when this happened, subjects would go down fruitless search paths due to the early error. The time results give some objective evidence towards this nature of this penalty.

Ubiquitous recording intrinsically has privacy implications. For the present evaluation, conventional safeguards ensured compliance (e.g., informed consent, awareness, deletion features, etc.) and data protection (e.g., password, firewall, encryption, etc.). For a real-world deployment, reasonable protection is possible, but compliance is harder. Improved awareness and control mechanisms for recordees are needed; education of the heterogeneous, worldwide, social customs and laws on recording are sensible. The social awkwardness of repeatedly asking for permission to record conversations affected the RA and may continue to be a good inhibitor, for now.

Ubiquitous computing enables the Memex vision of long-term personal data archival of everyday experiences. Our experience with iRemember represents early evidence that such archival can serve as a valuable resource for memory problems.

7. REFERENCES

[1] Abowd, G.D. Classroom 2000: An Experiment with the Instrumentation of a Living Educational Environment. *IBM Systems Journal*, **38**(4), 508–530, (1999).

[2] Bush, V. As We May Think. *Atlantic Monthly* **76**(1), 101–108. (July 1945).

[3] Eldridge M., Sellen A., and Bekerian D., Memory Problems at Work: Their Range, Frequency, and Severity. Technical Report EPC-1992-129. Rank Xerox Research Centre. (1992).

[4] Gemmell, J., Bell, G., Lueder, R., Drucker, S., and Wong, C., MyLifeBits: Fulfilling the Memex Vision, Proc. *ACM Multimedia '02*, Juan-les-Pins, France, 235–238. (2002).

[5] Kletz, T. Lessons from Disaster: How Organizations Have No Memory and Accidents Recur. Institution of Chemical Engineers. Rugby, Warwickshire, UK. (1993).

[6] Lamming, M. and Flynn, M. "Forget-me-not- Intimate Computing in Support of Human Memory. In *Proceedings of FRIEND21, Intl. Symposium on Next Generation Human Interface*, Megufo Gajoen, Japan (1994).

[7] LifeBlog, Nokia, http://www.nokia.com/lifeblog

[8] Linton, M. "Memory for real-world events." In Norman, D.A. and Rumelhart, D.E. (eds.), *Explorations in cognition* (Chapter 14). San Francisco: Freeman. (1975).

[9] Loftus, E.F. *Eyewitness Testimony*. Harvard Univ. Press, Cambridge, Massachusetts, (1996).

[10] Moran, T.P., Palen, L., Harrison, S., Chiu, P., Kimber, D., Minneman, S., van Melle, W., and Zellweger, P. "I'll get that off the audio": A case study of salvaging multimediameeting records. *Proc. of CHI'97*. (1997).

[11] Rhodes, B. *Just-In-Time Information Retrieval*. Ph.D. Dissertation, MIT Media Lab (May 2000).

[12] Schacter, D.L. The Seven Sins of Memory: Insights from Psychology and Cognitive Neuroscience. *American Psychologist*. **54**(3), 182–203 (1999).

[13] Stifelman, L., Arons, B., and Schmandt, C. The audio notebook: paper and pen interaction with structured speech. In *Proceedings of the SIG-CHI on Human factors in computing systems*. 182–189. (2001).

[14] Vemuri, S. *Personal, long-term memory aids*. Ph.D. Dissertation, MIT Media Lab (February 2005).

[15] Vemuri, S., Bender, W., Next-generation personal memory aids. In *BT Technology Journal*. **22**(4) 125–138 (October 2004).

[16] Vemuri, S., Schmandt, C., Bender, W. An Audio-Based Personal Memory Aid. Proc. *Ubicomp 2004*.

[17] ViaVoice, http://www-3.ibm.com/software/speech/

[18] Wagenaar, W.A. My Memory: A study of Autobiographical Memory over Six Years. In *Cognitive Psychology*. **18**, 225–52 (1986).

[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. *Proc. SIGIR99*. 26–33. (1999).

[20] Wong, B.A., Starner, T.E., and McGuire, R.M. Towards Conversational Speech Recognition for a Wearable Computer Based Appointment Scheduling Agent. GVU Tech Report GIT-GVU-02-17. (2002).