

# An Audio-Based Personal Memory Aid

Sunil Vemuri, Chris Schmandt, Walter Bender, Stefanie Tellex, Brad Lassey

MIT Media Lab  
20 Ames St.  
Cambridge, MA 02139 USA  
{vemuri,geek,walter,stefie10,lassey}@media.mit.edu

**Abstract.** We are developing a wearable device that attempts to alleviate some everyday memory problems. The “memory prosthesis” records audio and contextual information from conversations and provides a suite of retrieval tools (on both the wearable and a personal computer) to help users access forgotten memories in a timely fashion. This paper describes the wearable device, the personal-computer-based retrieval tool, and their supporting technologies. Anecdotal observations based on real-world use and quantitative results based on a controlled memory-retrieval task are reported. Finally, some social, legal, and design challenges of ubiquitous recording and remembering via a personal audio archive are discussed.

## 1 Introduction

The idea of recording everything in one’s life is not new. An early proposal, Memex, dates back to 1945 [4]. Just over a decade ago, ubiquitous audio recording systems were built and studied using desktop computers and workstations [12]. Now, improved portability facilitates increased ubiquity and it follows that more and more industry, academic, and government groups are investigating the possibility of recording everything in one’s life [8,15,16]. Indeed, computers have reached the point at which continuous, verbatim recording of an individual’s life experiences is technologically and economically feasible. The challenges of miniaturization, battery life, storage capacity, and aesthetics are being addressed. A few years ago, palm-sized PDAs with high-speed, wireless networking were choice platforms for these applications; now, miniature units (e.g., iPods) are available with recording ability and capacity to store weeks, if not months, of data. It may be sooner than we think before a commercially-available watch-sized device arrives capable of ubiquitous recording. But recording everything is the easy part.

The outstanding challenge is turning vast repositories of personal recordings into a useful resource while respecting the social, legal, and ethical ramifications of ubiquitous recording. This paper examines amassing data for the purpose of helping people remedy common, everyday memory problems. Previous attempts at building computational, portable memory aids have studied note-taking [22] and passive recording of proximity to people, office devices, and other environmental markers [14]. Memory

studies have examined long-term recall in real-world settings [30] and audio recall in laboratory settings [33], but the combination of audio and long-term memory in real-world situations is relatively unexplored.

A sampling of related efforts involved in collecting digital audio and video archives include classrooms settings [1], meetings [20], voicemail [32], workplace and telephone discussions [12], informal situations [16], and personal histories [8]. Among these and other similar projects in the domain, indexing techniques such as large-vocabulary speech recognition, speaker identification, face recognition, and audio/video scene analysis are used to mine noisy data for search cues. Whereas much attention has been given to improving the accuracy and robustness of these techniques, less has been done on designing tools to help users make sense of noisy data sources and to understand which analyses prove most useful to specific tasks. For example, differences are expected among the search strategies employed by those trying to remember a previously-witnessed event versus someone trying to find information within an unfamiliar collection; in the former case, retrieving any information that triggers the memory of the event is sufficient; in the later, finding the exact information is necessary.

This paper examines these techniques in light of alleviating memory problems. To this end, we have constructed a wearable “memory prosthesis.” The device records audio from conversations (and other sources), applies large-vocabulary speech recognition to it, and provides a suite of retrieval tools to help the user access forgotten memories. Experiences of one of the authors, who has used the device for recording a variety of everyday experiences for two years, are reported. To better understand the utility of such voluminous collections of recordings for memory assistance, and which among a wide range of interaction techniques should be immediately on-hand when a memory problem occurs, an experiment was conducted with an exemplar memory-retrieval task. Before describing the technology and experiment, some relevant background on memory in general is provided. The paper concludes with a discussion of some social and legal implications of ubiquitous recording.

## 1.1 Memory

Schacter’s taxonomy, or the “Seven Deadly Sins of Memory,” succinctly describes the most common memory problems [24]. The six involving forgetting and distortion are shown in Table 1. The seventh, “persistence” (i.e., pathological inability to forget), is of less interest to memory-aid designers.

**Table 1:** Six of the seven “Sins of Memory” [24]

| <b>Forgetting</b>   | <b>Distortion</b>                                       |
|---|---|
| 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)           |

It is not the goal of the present research nor is it expected that a single memory aid can adequately address all such problems; instead, the focus is to address a subset. Previous studies [6] explore the frequency of some types of forgetting in workplace settings (Table 2). In Schacter’s taxonomy, “retrospective memories” are analogous to “transience”; “prospective memory” and “action slips” are both forms of “absent-mindedness.”

**Table 2:** Frequency of some common memory problems in the workplace [6]

| Type                 |     | Description   | Example   |
|----------------------|-----|---|---|
| Retrospective Memory | 47% | Remembering past events or information acquired in the past                                     | Forgetting someone’s name, a word, an item on a list, a past event          |
| Prospective Memory   | 29% | Failure to remember to do something   | Forgetting to send a letter, forgetting an appointment                      |
| Action Slips         | 24% | Very short-term memory failures that cause problems for the actions currently being carried out | Forgetting to check the motor oil level in the car before leaving on a trip |

The prototype aims to address transience, the most frequent among the memory problems. 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 or needs to remedy the problem, what is needed is a resource to help. This is where the memory prosthesis comes into play.

## 1.2 Design Goals

The approach to address transience and blocking memory problems is to build capture tools to record daily experiences and retrieval tools to find memory triggers that remedy these problems. Although the current research prototypes do not perfectly achieve all of the ideals described below, they are sufficient to allow sympathetic subjects to start experiencing portable, ubiquitous recording and are proving instructive towards the initial design of memory retrieval tools and validating the approach.

### Data Capture

One of the early questions in the design of the memory prosthesis is what data sources should be captured. An ideal data source has maximal memory-triggering value while presenting minimal computational and storage demands. Furthermore, a design goal is to minimize the effort needed by the user to capture daily experiences. This means, when possible, data should be captured with little effort or passively. Finally, to minimize the chance of missing a potentially-valuable memory trigger (at the cost of retaining superfluous ones), nearly-continuous recording of daily activity is desired. To these ends, a wearable recording apparatus was constructed (Section 2).

High on the list of desired data sources was audio, due to the anticipated memory-triggering value of verbatim audio, the desire to capture conversations that occurred in informal settings, the ease of capturing audio using a wearable device, the tractable data-storage requirements, and the readily-available content-analysis tools (e.g., speech recognition). However, for legal and human-subject approval reasons, audio recording requires consent from all participants for each recording. Consequently, collecting these data can neither be completely passive nor continuous. Similar to doppelgänger [21], sources that are captured and archived both continuously and passively include the user's location, calendar, email, commonly-visited web sites, and weather.

Video was considered, but the hardware complexity of a wearable, continuous-video-capture device combined with the data storage requirements and difficulty of analyzing and indexing hundreds of hours of video suggested otherwise. Even so, research on video-retrieval systems has gained momentum with some promising results [26]. Still photography is a common request and is being integrated. Capturing biometrics [9] was also considered and remains an item for future work.

It should be noted that completely passive data capture may in fact hurt memory recollection as evidenced by the disadvantage of no note-taking among students in classroom situations [18]. The choice to prefer passive data capture is to simplify and reduce forgetfulness in the daily data-capture process. Ironically, forgetting to activate the memory prosthesis (i.e., absent-mindedness) is a common problem.

## **Retrieval**

An ideal memory aid would proactively determine when a memory problem occurs and provide just-in-time minimally-intrusive remedies. We are far from this goal and the present approach and technologies still requires active effort.

With enough time and incentive, users could probably remedy most memory problems via a comprehensive archive of past experiences. Since most real-world situations do not provide either such time or incentive, the primary goal for the memory retrieval tools is to remedy problems within the constraints of attention and effort users can and are willing to commit. There is still limited empirical evidence about such willingness: Section 4.1 provides some anecdotal evidence about real-world use. The evaluation described in Section 4.2 offers some insights about memory problem frequencies, how subjects approach memory repair, and what technologies are most useful when given an investigator-invented memory-retrieval task.

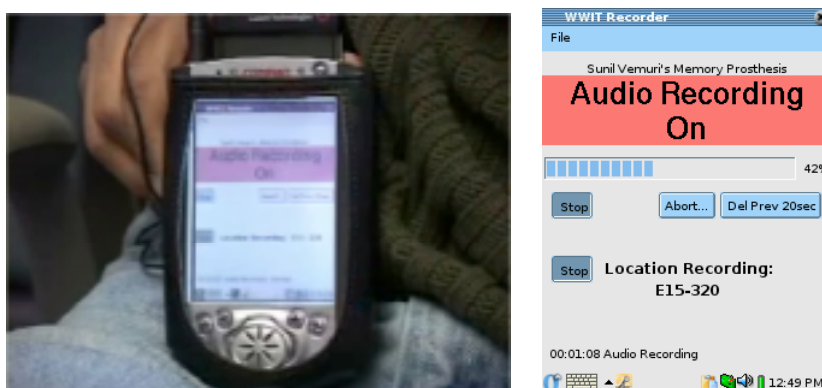
Given the present focus on the workplace settings, it is anticipated that users could benefit from retrieval tools on both handheld/wearable devices and desktop/laptop personal computers. Hence, a goal was to design for all of these platforms, taking advantage of the idiosyncrasies of each. This paper describes the wearable capture device (Section 2) and the personal-computer-based retrieval tool (Section 3).

One of the challenges to the design of the prototype is finding ways to manage the large quantity of collected data that is often rife with both noise due to the intrinsic limitations of computational information-extraction from the selected data sources and irrelevance with respect to a specific information need. The former is more typical of a personal memory assistant and will be addressed in more detail throughout the paper. The latter point is addressed in a manner similar to conventional text corpora searches: keyword searches with ranked retrieval results.

## 2 Memory Prosthesis

The wearable capture device (Figure 1) is implemented on an iPaq PDA coupled with an 802.11b wireless network card. When activated by the user, the device records audio and physical location. It determines location via proximity to numerous stationary 802.11b base stations primarily inside the building. Recorded data are streamed to a server and access is restricted to within our local computer network due to the data-security requests of the users. Previously, the device had been in use outside the local network, including nearby publicly-accessible wireless-networks. The server also passively captures the additional sources mentioned in Section 1.2.

The design reflects a variety of considerations necessary to balance the requirements of obtaining speech-recognition-quality audio while meeting necessary social and legal requirements. Moreover, the investigators tried to satisfy a higher standard by ensuring co-workers felt reasonably comfortable with the presence of potentially privacy-invading technology. The device includes a display with the audio recording status, an audio-level meter, and recordees may request recording deletion via anonymized email.



**Figure 1:** Prototype “Memory Prosthesis” (left) with close-up of screen on the right

When recording, users often move the device from its waist-level belt-clip in favor of chest height. This allows conversation partners a clear view of the screen, the recording status of the device, and serves as a constant reminder that the conversation is being recorded. Chest-height positioning also provides better acoustic conditions for the built-in near-field microphone on the iPaq. An attachable lavalier microphone is also occasionally used. Although speech-recognition-quality recording of all participants in a conversation is desired, this is not always feasible since no cost-effective, portable recording solutions could be found. Placing the iPaq halfway between two speakers results in bad audio for both. Placing the device at chest-height or using the lavalier allows at least one of the speakers, the wearer, to be recorded at adequate quality at the cost of poor quality recordings for all other speakers.


One might question the earlier assertion that “recording is the easy part.” Given the forthcoming descriptions of memory-retrieval challenges and the social and legal ramifications, it most certainly remains that “recording is the *easier* part.”

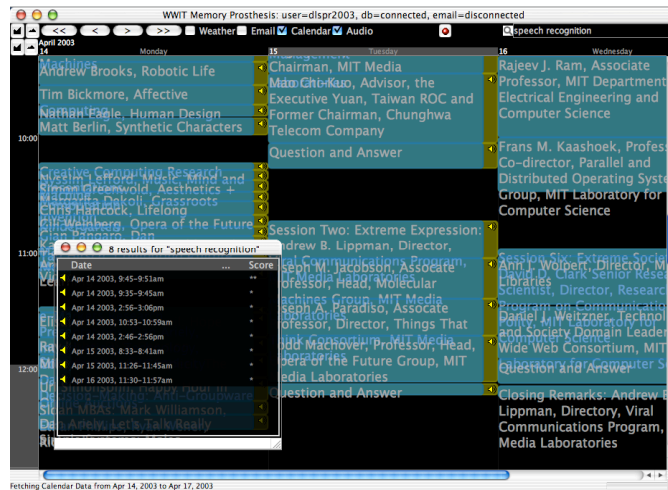
### 3 Memory Retrieval

An anticipated consequence of a daily-worn recording device is the accrual of years of personal interaction and conversation data. To help alleviate transience and blocking memory problems, it is expected that users would want to search such a collection for memory triggers. Given the focus on audio and the state-of-the-art in information retrieval technologies, capabilities such as browsing, skimming, and free-form keyword searches of timeline-organized audio archives were selected for implementation.

We implemented the first memory retrieval tool for use on a personal computer due to the expected desire for retrieval on this platform in workplace settings and the ease of implementing features (some of which are computationally and graphically demanding) compared to a PDA. Little is known about memory retrieval using personal data archives. Hence, one goal of the evaluation described in Section 4.2 is to begin understanding how users approach memory repair and which among the many features prove worthy of inclusion in subsequent iterations and implementation on the PDA.

#### 3.1 Searching Audio Collections

Figure 2 shows the personal-computer-based interface available to search and browse large collections of recordings. Each column along the x-axis corresponds to a single day; time-of-day is on the y-axis. Each calendar event has a corresponding blue rectangle. Audio recordings are represented with a yellow rectangle and an associated icon . Zooming and panning features are provided; double-clicking on an item opens a more detailed view of a recording. Color transparency is used to blend overlapping elements. Cluttering of text from neighboring items is possible and individual items can be made more legible by clicking on them.



**Figure 2:** Visual interface for browsing and searching through recordings. A keyword-search feature (upper right) is available with example results shown in lower left.

Error-laden transcripts are generated using IBM's ViaVoice [29] speech-recognition software and users can perform relevance-ranked keyword searches on these transcripts using the Lucene search engine [17]. Issues related to keyword-searching on speech-recognized text are covered in Section 5.3.

The data in Figure 2 reflect the agenda and recordings that took place as part of a three-day conference. This is the same set of conference recordings used in the study described in Section 4.2. For this data set, the calendar entries were copied verbatim from the conference agenda. Word error rate (WER) for speech-recognition systems is defined as the sum of insertion, deletion, and substitution errors divided by the number of words in the perfect transcript. Though not formally computed for the entire set, representative samples show a uniform WER distribution between 30–75%. Variation seemed to depend on speaker clarity, speaking rate, and accent.

### 3.2 Finding Memory Triggers in a Single Recording

Influenced by ScanMail [32], the interface described in this section addresses the problem of finding memory triggers within a single recording (Figure 3). Listening to long recordings is tedious and browsing error-laden transcripts is challenging [27]. Since recordings may last hours, the interface attempts to: (1) help the user find keywords in error-laden transcripts; (2) bias the user's attention towards higher quality audio; (3) help the user recall the gist of the recording; and (4) provide ways to play audio summaries that may serve as good memory triggers.

Several features are included to improve the utility of the speech-recognizer-generated transcripts. A juncture-pause-detection algorithm is used to separate text into paragraphs. Next, similar to the Intelligent Ear [25], the transcript is rendered with each word's brightness corresponding to its speech-recognizer-reported confidence. A "brightness threshold" slider allows users to dim words whose confidence is below the threshold. This can be used to focus attention on words that are more likely to be recognized correctly. Next, an option to dim all English-language stopwords (i.e., very common words like "a", "an", "the", etc.) allows users to focus only on keywords. A rudimentary speaker-identification algorithm was included and the identifications are reflected as different colored text (seen as red or aqua in Figure 3).

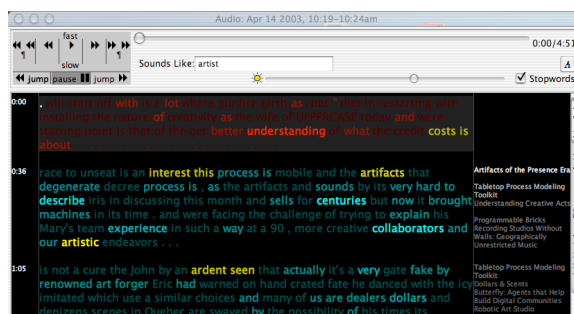


Figure 3: Interface for browsing, searching, and playing an individual recording

A phonetic or “sounds-like” search feature is provided to help users identify misrecognized, out-of-vocabulary, and morphological-variant words by locating phoneme-sequence similarity (seen as yellow text and as dark blue marks in the scrollbar in Figure 3). The phonetic search algorithm is similar to [31] and uses CMUDict [5].

A sophisticated audio-playback controller (based on SpeechSkimmer [2]) capable of audio skimming is provided (upper left of Figure 3). Depending on where the user clicks in the controller, audio plays at normal, fast, or slow rates (non-pitch adjusted). The control offers two forms of forward and reverse play: one skips periods of silence and the second plays only the first five seconds of each paragraph.

Finally, annotations to the transcript (seen as white text on the right of Figure 3) show information potentially related to the neighboring transcript text. To generate this, each section of the transcript text is used as a query to a search engine serving a corpus of roughly 800 past and present department-related project abstracts. The top-ranked project names are displayed.

It should be noted that in addition to the features described in this section, the software currently in use by one of the authors also has the ability to include the other data sources mentioned in Section 1.2 (i.e., email, location, weather, web sites) as part of the memory retrieval process. These tools include additional visualization and search features tuned to these data sources.

Again, it was not clear which, if any, among the retrieval capabilities in either interfaces shown in Figures 2 and 3 would prove useful for memory assistance. Among other issues, the study described in Section 4.2 was designed to explore this.

## 4 Evaluation

The evaluations presented in this section include both anecdotal reports from two subjects (one of whom is an author) who use the memory prosthesis for real-world recording tasks (Section 4.1) and a controlled study of the memory-retrieval tool via artificially-induced memory problems among subjects who are not memory prosthesis users (Section 4.2). The purpose of these evaluations respectively is to (1) identify what tasks people choose to use portable, ubiquitous recording devices; and (2) confirm the basic assumption that an audio-based information-retrieval tool can serve as a memory aid for blocking and transience memory problems.

Short of a longitudinal large-scale deployment, it is unlikely that an adequately-controlled evaluation over a general populace can be performed. The present approach attempts to approximate the likely requirements of a memory prosthesis user. Examples of more controllable, memory-demanding tasks include students searching through archives of recorded lectures in preparation for a test, or conference attendees wishing to write trip reports. Speech from such presentation-style settings is obviously different from conversations and participants are likely to be less-engaged in a such settings. While neither scenario is ideal, conference talks were chosen, and these studies have helped inform subsequent iterations of the memory retrieval tools.



#### 4.1 Experiences with Ubiquitous Recording

Given the opportunity to record anytime, anywhere, what would one record? Who would one record? Are the recordings useful to a real-world task? In addition to the author (Student P), another student (Student Q) in the department (not directly involved in the research project) is also a frequent audio-recorder of select conversations. Student Q uses a variety of computer-based audio recording tools, including the memory prosthesis. Experiences with the former and interviews with the latter are reported.

Student P has a vested interest in frequent recording and chooses to record at most opportunities (within the bounds of legal, social, and human-subject protocol guidelines). Most of those recorded are within the student's circle of trusted co-workers and have *never* expressed hesitation about recording work-related matters. Others within the circle have been asked and have agreed to be recorded, but Student P felt a sense of recordee-discomfort (either by overt statements, tone of voice, or body language). When among these co-workers, Student P is more cautious about the circumstances and topics when asking permission to record and errs on the side of not recording.

Student Q opts to record only a limited set of experiences: student-advisor conversations. In particular, conversations focused around an impending, time-critical task such as writing a paper: "The things that I record are things that I think ahead of time will be critical and I will need in a short amount of time. I don't record anything or everything...What I do make a point of recording is conversations with my advisor." This recording strategy reflects a desire to minimize absent-mindedness problems in contrast to Student P who wishes to reduce transience.

Similar to Moran et al.'s observations [19], Student Q uses audio recording as a backup to hand-written note-taking: "If I write things down, it's... like pointers into the recording and it organizes the thoughts for me." With regard to recording usage, "I often don't listen to the recording. But, to know that I have it is good. Sometimes I can't read my own writing. If I manage to write enough... that's fine. Sometimes I don't. So then I go and see what I missed, or see what I was trying to write down."

Interestingly, Student P, who has access to a much larger collection of recordings, mainly uses the recordings for the identical real-world task: retrieving student-advisor discussions in preparation for a pending writing deadline. Both students cite common reasons for recording and usage including: limited advisor availability, the anticipated high quality of the conversations, and time-pressure of the task.

Student Q—when using a non-memory-prosthesis recorder—opts not to archive recordings for space-saving reasons, but expresses no objection to keeping them. The recordings are often discarded after the task deadline, and anything important is transcribed before disposal.

#### 4.2 Evaluation of the memory-retrieval aid

To better understand the memory-retrieval tool's value among a larger and broader population, investigators confronted non-memory-prosthesis subjects with artificially-induced memory problems based on past-witnessed events and asked them to resolve these with the tool. Specifically, subjects were presented with a remembering-and-

finding task based on information presented in a three-day conference that occurred approximately one month prior to the test. Some subjects were speakers.

Subjects were given a questionnaire with 27 questions. Investigators designed questions with varying difficulty, some with multiple parts, and whose answers could be found and answered unambiguously. For example, a relatively simple question is “What book (title and author) does [Speaker X] cite?” and a more difficult, time-consuming example is “Which projects or people specifically cite the use of a speech recognizer?” Some questions could be answered simply by performing searches using words found in the question, some could not be answered unless the subject remembered the answer or was able to find the answer in the audio.

To maximize audio quality, speakers were recorded using the auditorium’s existing recording apparatus and the audio was fed into the memory prosthesis system. This approach, as opposed to using the memory prosthesis device, was in anticipation of ubiquitous-computing-friendly environments that can readily support such data transfers. In total, approximately 14 hours of audio was available from 59 talks. Talk lengths ranged from two to 45 minutes.

No subject attended the entire conference, but all attended at least part of it. The question dispersion was intended to ensure each subject witnessed some of the answers. It was hoped, but not guaranteed, that each subject would find and attempt questions in which they remembered witnessing the answer at the event (as opposed to knowing the answer via another source), and some memory problem occurred such that they required assistance to answer. Hence, the three experimental conditions were:

- C1: Unaided: Subjects answered questions without any assistance, either by remembering the answer from the event or from another source.
- C2: Aided, Witnessed: Subjects answered a question using both the memory aid and information they remembered from the event.
- C3: Aided, Non-witness: Subjects did not witness the event and their answer was based on examination of data during the experiment and possibly using their previous knowledge of the department, its people, and their research in general.

Before the questions were administered, subjects were given a 5–10 minute training session with the memory retrieval software (described in Section 3). After training, subjects were given the full questionnaire. The test was split into two phases. In Phase 1, subjects were asked to answer any questions they already knew, but without the use of any aids (Condition C1). Subjects were instructed to answer as many or as few questions as they wished and were given as much time as needed to answer these questions. In addition to collecting C1 data, this phase allowed subjects to familiarize themselves with all of the questions without any time pressure in preparation for Phase 2. A pilot study indicated that subjects did not read all the questions if under time-pressure. All subjects completed Phase 1 in less than 15 minutes.

After a subject finished unassisted question-answering, Phase 2 began in which the memory-retrieval software tool was provided. However, subjects were now limited to 15 minutes to answer any remaining questions. The reasons for having a time limit included: (1) encouraging subjects to prioritize the questions they wanted to attempt;

and (2) putting a limit on a subject's time commitment. Subjects were allowed to ask user-interface clarification questions without time penalty.

During both phases, subjects were asked to verbalize their thought process as they answered questions. This was audio recorded and an investigator remained in the room taking notes. Upon completion, subjects were informed about the nature of the experiment (i.e. studying search strategies), interviewed about their experience, asked to reflect on their thought process, and elaborate on any details that they might not have mentioned during the task.

Hypotheses are listed below. Time is computed on a per-question basis. Success rate is the number of questions answered correctly or partially-correctly divided by the number of attempts. In addition to these hypotheses, the investigators are also interested in what strategies subjects employed when remedying memory problems and what user interface features subjects found most useful.

### **Hypotheses**

H1a: Unaided question-answering (C1) will take less time than aided (C2 & C3).

H1b: Unaided question-answering (C1) will have a lower success rate compared to Condition C2 (aided, witnessed).

H2a: Aided question-answering of previously-witnessed events (C2) will take less time than aided question-answering of non-witnessed events (C3).

H2b: Aided question-answering of previously-witnessed events (C2) will have a higher success rate compared to aided question-answering of non-witnessed events (C3).

## **5 Results**

Subjects included three women and eight men ranging in age from 18 to 50. Nine were speakers at the conference; two were non-native English speakers. Subjects' prior exposure to the research presented at the conference ranged from one month to many years and in varying capacities (e.g., students, faculty, visitors). No subject attended the entire conference. Among the speakers, most attended only the session in which they spoke. Nine subjects claimed to be fatigued during the conference and all said they were occupied with other activities (email, web browsing, chatting, daydreaming, preparing for talk, etc.) at least part of the time. Investigators classified six subjects who have prior understanding of how speech-recognition technology works.

### **5.1 Phase 1 (Unaided Memory)**

Answers were labeled as one of "correct," "partially correct," "incorrect," or "no answer." A "correct" answer satisfies all aspects of the question correctly. A "partially correct" answer has at least one aspect of the question correct but another part either incorrect, omitted, or includes erroneous extraneous information. This labeling was common for multi-part questions. An "incorrect" answer has no aspects correct. A "no

answer” is one in which subjects attempted to answer, verbalizations indicated a memory problem, but no answer was submitted. Among all of these, subjects spent on average 29 seconds to read, verbalize, and answer (or choose to not-answer) a question.

Memory problems that occurred were noted. If investigators either observed or a subject verbalized a memory problem while answering, investigators classified it as one of the six problems listed in Table 1. If it was clear that a memory problem occurred, but there was ambiguity between two types of memory problems, each was assigned 0.5 points. If it was not clear if a memory problem occurred, no points were assigned. In some cases, subjects misread the question, and consequently, answered incorrectly. These were not counted. Aggregate results are summarized in Table 3. Investigators did not observe any of the following memory problems: absent-mindedness, bias, suggestibility, or persistence.

These results correspond to the C1 condition. Without a basis for comparison, it is difficult to say whether the question-answering performances are good or bad. Regardless, the interest with respect to the memory aid designer is the types and frequencies of memory problems. In the present case, both transience and blocking problems were found as expected, but misattribution problems were unexpectedly common. Phase 2 examines how subjects approached the task of remedying some of these.

**Table 3:** Phase 1 question-answering tallies and memory problem categorization

| Answer            |    | Problem        |    |
|-------------------|----|----------------|----|
| Correct           | 47 | Transience     | 22 |
| Partially Correct | 27 | Blocking       | 4  |
| Incorrect         | 20 | Misattribution | 9  |
| No Answer         | 12 |                |    |

## 5.2 Phase 2 (Aided Memory)

As mentioned previously, the hope in Phase 2 is to observe subjects attempting to remedy memory problems (Condition C2) and understand the strategies they employ under that condition. In Phase 2, all subjects engaged in a series of searches. A search attempt begins with the subject reading the question and ends with them selecting another question (or time runs out). The memory retrieval software recorded logs of what the subject was doing and which question they were answering. This was used in conjunction with the audio recordings of the experimental session to determine time spent on each question. Investigators further classified each question-answering attempt as either Condition C2 (subject witnessed the answer at the original event) or C3 (subject did not witness the answer). Classification was based on subject verbalizations and post-experiment interviews. Finally, attempts were classified as “successful” if the subject correctly or partially-correctly answered a new question, verified a previously-answered question, or improved an answer to a previously-answered question. An attempt was classified as “no answer” if the subject gave up on the search without producing an answer or time ran out. In no instances did a subject provide an incorrect answer during Phase 2. Results are detailed in Table 4 and summarized in Table 5.

The 82% success rate under C2 versus the 70% rate under C1 gives some support for Hypothesis H1b (i.e., higher success rate when aided). While also not conclusive, there is some support for Hypothesis H2b: subjects were able to answer questions more successfully in C2 (82%) compared to C3 (53%). Not surprisingly, in support of Hypothesis H1a, time spent per question under C1 was less than both C2 and C3 ( $p < 0.0001$ ). However, with no statistically-significant mean differences between C2 and C3 timing, there is no support for Hypothesis H2a.

**Table 4:** Phase 2 results. Each subject's (A–K) answering-attempts shown in sequence with time spent in subscript. 🕒 = subject witnessed answer (no icon=non-witness); ✓ = successful attempt (no icon=no answer). Time ran out during the last entry on each row.

|   |          |          |          |        |        |        |        |      |      |      |      |  |
|---|----------|----------|----------|--------|--------|--------|--------|------|------|------|------|--|
| A | 3:22     | ✓ 0:56   | ✓ 3:40   | 3:30   | 3:30   |        |        |      |      |      |      |  |
| B | 🕒 7:40   | 🕒 ✓ 5:31 | 1:44     |        |        |        |        |      |      |      |      |  |
| C | 🕒 ✓ 2:04 | 0:54     | ✓ 1:49   | ✓ 1:31 | ✓ 3:12 | ✓ 1:55 | ✓ 1:34 | 0:39 | 1:20 | 1:10 | 2:09 |  |
| D | 🕒 ✓ 2:44 | 5:07     | 3:02     | ✓ 4:59 |        |        |        |      |      |      |      |  |
| E | 🕒 ✓ 1:40 | 1:12     | ✓ 3:00   | 1:47   | 3:02   | 4:18   |        |      |      |      |      |  |
| F | ✓ 3:19   | ✓ 2:23   | ✓ 2:14   | 2:05   | 2:33   | 2:10   |        |      |      |      |      |  |
| G | 🕒 ✓ 2:39 | ✓ 1:46   | 7:11     | 3:51   |        |        |        |      |      |      |      |  |
| H | 🕒 ✓ 2:32 | 🕒 ✓ 2:40 | ✓ 3:48   | ✓ 3:25 | 3:07   |        |        |      |      |      |      |  |
| I | 1:00     | ✓ 4:12   | 🕒 ✓ 1:48 | 1:57   | ✓ 1:38 | ✓ 2:01 | ✓ 2:48 |      |      |      |      |  |
| J | 2:57     | 🕒 ✓ 1:37 | ✓ 3:52   | 3:03   | 1:14   | 2:48   | 1:06   |      |      |      |      |  |
| K | ✓ 2:01   | ✓ 1:21   | 3:22     | 🕒 4:06 | ✓ 1:50 | 2:25   |        |      |      |      |      |  |

**Table 5:** Summary of Table 4 question-answering times (in seconds) and question-answering tallies (not counting C3, no answer timeouts)

|           | Witness (C2) 🕒 |         | Non-Witness (C3) |         |
|-----------|----------------|---------|------------------|---------|
|           | Success ✓      | No Ans. | Success ✓        | No Ans. |
| Mean      | 155            | 353     | 160              | 154     |
| Std. Dev. | 71             | 151     | 73               | 94      |
| N         | 9              | 2       | 23               | 20      |

The timing data in general has caveats. Some subjects found the interface initially challenging, yet learned how to use it better over time: “I think I'm getting better at figuring out how to search the audio just in terms of thinking about things that might work.” Furthermore, question difficulty was not uniform and not all subjects formulated an optimal strategy to maximize the number of answers solved. For example, some subjects intentionally chose questions out of curiosity versus optimizing their overall task-performance. Such confounding factors make it difficult to draw conclusions on timing differences. However, the timing similarity between C2 and C3 might suggest subjects have a condition-independent time threshold (roughly 4 minutes) after which they will move on whether they find an answer or not.

Observations during the experiment revealed what aspects of the interfaces (Figures 2 and 3) were most valuable. Among the nine instances in which subjects were able to remedy failures, seven initiated searches by correctly identifying the talk in the calendar; the remaining two found the correct talk by keyword searching. Once in the recording, in six instances, subjects used phoneme searching to identify potentially-relevant sections and limited audio playback to those. In two instances, subjects played the audio from the beginning until the answer was heard, and in one instance, the subject skipped to various points in the recording, using the transcript as a guide, until finding the answer. In one instance in which a subject was a witness but failed to find an answer, a misattribution memory problem occurred causing the subject to initially open the wrong recording. After four minutes of futile searching within the wrong audio clip, the subject gave up. In the other instance, the subject initially found the right recording, tried listening to audio from various sections (using the transcript as a guide) and phonetic searching, but to no avail.

### 5.3 Searching via Speech Recognition

Keyword-searching of audio collections is problematic due to the inherent errors in speech-recognizer-generated transcripts. Not surprisingly, subjects stated that poor-quality speech recognition made the task challenging. Optimistic predictions aside, high-accuracy speech recognition of conversations in poorly-microphoned, heterogeneous environments will not happen anytime soon. Despite this, high-quality transcription—while beneficial—may not be necessary, especially for memory-retrieval tasks.

Witbrock [34] suggests general-purpose audio-information retrieval tasks can still be performed at high WER. Speech recognition has been shown to help in voicemail-retrieval [32] and calendar-scheduling tasks [35]. In previous studies, we found that error-laden speech-recognizer-generated transcripts synchronized with time-compressed audio playback, can improve subject comprehension, especially when word-brightness is rendered proportional to recognizer-reported confidence [28]. Techniques to build and evaluate information-retrieval systems for broadcast-news recorded-speech collections have been studied in detail as part of the TREC Spoken Document Retrieval (SDR) task [7].

In the present study, previous experience with speech recognition seemed to be useful. For example, one subject typed the query “brazil” to find “Breazeal” since the latter was expected to be out-of-vocabulary. Another subject focused on questions that included keywords suspected of being in the recognizer’s vocabulary. Other subjects commented that an adjustment period is needed to learn the novel search features and peculiar limitations of searching speech-recognizer-generated transcripts. These subjects added that their adjustment seemed to begin within the 15-minute testing period.

The phonetic “sounds like” searching feature was used often in Phase 2. However, the out-of-vocabulary problem was still observed when subjects attempted queries with domain-specific keywords such as “BeatBugs” and “OMCSNet.” The absence of these words from CMUDict [5] prevents a phonetic translation. The overall sense from subject feedback was that this was useful despite the limitation.

## 6 Discussion

The results give the authors both relief and confidence that the current memory retrieval aid is a good starting point for memory retrieval via audio-based search. Subjects found answers within large collections of audio recordings, typically within a few minutes. In most cases, subjects were able to strategize ways to use the tool along with their remembrance of the past to identify the correct recording and to localize within a recording to the answer. Speech recognition, despite poor-quality transcripts, was useful for both keyword-searching audio collections and for helping subjects localize and select sections of recordings to play back.

Lessons learned from the personal computer study are being applied to our handheld/wearable solution. Searching the calendar and audio playback were common in the witness condition; these capabilities have transferred well to the PDA. Current PDAs lack the computational resources needed for large-vocabulary speech recognition and real-time phonetic searching. Thus, a server-based approach is still required, but this will likely change within a few years. PDA screen-size issues remain as subjects often found the large personal-computer display useful to visually skim large sections of transcripts to help them localize their audio playback choices.

Text-based keyword-searching may not transfer easily to a PDA or wearable device due to text-entry limitations. Query input via speech is an intriguing possibility and speech-recognizer-mediated input for a more-constrained wearable calendaring system has been studied [35]. High WER and out-of-vocabulary problems suggests this approach may not be well suited for exact-match queries. Phonetic searching, with its looser constraints, may be better suited and might even produce better results. On today's typical PDA, it is possible to generate phonemic transcripts, but text transcripts remain beyond current capabilities.

There is evidence suggesting that people remember better when the original context of the desired memory is reconstructed [10]. A PDA, used *in situ*, could take advantage of such context using, for example, an automatically-input-constrained search based on current location and other contextually-sensed factors.

The results, while emphasizing memory assistance, may have applicability to other audio-search domains. Subjects were able to find answers to questions from events that they did not witness, though not as accurately as in the witness-condition. Since subjects were familiar with the research in general, these results may not generalize to searching unfamiliar collections. But, there may be implications to organizational-memory applications. For example, absentees could have improved ways of quickly finding information within a recorded meeting. This example notwithstanding, memory fades over time and it is anticipated that the search process on events in the distant past will resemble that which is experienced by non-witnesses.

### Future Directions

Reported experience in the academic setting suggests one desirable scenario for ubiquitous recording: student-advisor task-based communications. Though a limited application area, it may generalize to other supervisee-supervisor communications, especially when task-based, limited-duration, high-quality, and infrequent. Further examination is

needed on this issue and a small-scale deployment of the memory prosthesis with non-investigators is underway.

What is not clear is the necessity for long-term archival for memory assistance. Task completion may be an opportune time for data purge, or at least, extraction of salient parts and deletion of the remainder. This limited- or non-archival strategy has additional benefits. First, a restricted search space may improve search experiences by reducing the time to find answers. Second, an illicit data intrusion would have limited ramifications. Third, if a user were embroiled in legal struggles in which recordings were subpoenaed, an established deletion strategy may avoid allegations of illegal destruction of evidence. Finally, conversation partners might be more willing to be recorded if there is an agreed-upon destruction policy.

## **7 Social and Legal Issues**

Even if memory prostheses and similar ubiquitous recording tools prove significantly valuable, they raise some obvious privacy concerns with respect to what permissions are needed to record, who will have access to the data, what social conventions are needed for such devices, and what legal protections are available. Common use is not expected until these points are adequately addressed.

Most states in the U.S.A. have laws requiring consent before initiating audio recording. There is some state-to-state variation with respect to both the setting (public versus private) and how many people must consent (one or all) [11]. While these standards describe what is legal, social conventions prescribe what is appropriate. The memory-prosthesis-wearing author has observed various reactions to the device. Consent is required and always requested prior to recording. Despite this, when the device is off, some conversation partners ask for verification before speaking freely, some assume the device is off. Also, the social greeting now includes a somewhat awkward request to be recorded (e.g., “Hello. Good to see you. May I record this?”). Others who have used the device found it uncomfortable to request such permission. This is especially true among those in supervisory roles who report more hesitation and discomfort when asking permission of their subordinates as compared to peers. Though people occasionally do decline requests to be recorded, it is not known if there are instances of accepted requests in which the person truly preferred not to be recorded, but agreed simply out of a sense of cooperation for a fellow researcher or other unspecified reasons. Brin posits ubiquitous recording is inevitable [3]. The experiences reported here suggest more studies are needed to understand what social conventions are appropriate for integration of personal recording devices into everyday life.

Assuming such conventions are possible, there is another reason for caution. Once recordings are made, there is very limited protection to prevent legal authorities from searching and seizing recordings via a court-approved warrant. The Fourth and Fifth Amendments to the U.S. Constitution describe protections against self-incrimination, search, and seizure. But, these protections are not expected to extend to memory prostheses. From a legal perspective, a close cousin to a memory prosthesis is a personal diary. US courts have addressed protection of personal diaries, and their current posi-



tion is diaries can be searched and seized [13]. Hence, it is unlikely that the less-private memory prosthesis would be afforded more protection.

Encrypting is an option. Hiding data is another option and the safest place might be inside one's body. While the courts have not set a limit to what can be seized, the standard for extracting things from one's body is higher than things outside [23].

## Conclusion

This paper presented a prototype, wearable memory aid with the goal of helping alleviate some everyday memory problems by creating a searchable, personal archive of everyday experiences. The prototype collects many data sources in support of this and the current focus is on audio.

A personal-computer-based memory retrieval tool allowing browsing, searching, and listening to audio and associated speech-recognizer-generated transcripts was presented. In contrast to traditional information-retrieval evaluations, the present study examines how the tool assists subjects who previously witnessed the recorded events and were familiar with the search collection. Results of a question-answering memory-retrieval task suggest that without assistance, mistakes are primarily attributed to memory problems such as transience, blocking, and misattribution. When their recollection was insufficient, subjects were able to use the retrieval tool in combination with bits they did remember to find answers. Similar to Whittaker et al.'s findings [32], the present observations also found error-laden speech-recognizer generated transcripts useful, especially when accompanied by the revised visualization and phonetic search features. Finally, some social and legal challenges associated with the ubiquitous recordings were presented.

## 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. Arons, B. SpeechSkimmer: Interactively Skimming Recorded Speech. *Proc. UIST 1993*, 187–196. (1993).
3. Brin, D. *Transparent Society*. Addison-Wesley (1998).
4. Bush, V. As We May Think. *Atlantic Monthly* **76**(1), 101–108. (July 1945).
5. CMU Pronouncing Dictionary. cmudict0.6d. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>
6. 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).
7. Garofolo, J., Auzanne, C., and Voorhees, E. The TREC Spoken Document Retrieval Track: A Success Story. *Proc. TREC 8*. 107–130. (1999).
8. 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).
9. Gerasimov, V., Selker, T., and Bender, W. Sensing and Effecting Environment with Extremity Computing Devices. *Offspring* **1**(1): 30–41. (2002).

10. Gooden, D., and Baddeley, A.D. When does context influence recognition memory? *British Journal of Psychology*. **71**, 99–104. (1980).
11. Hidden Cameras, Hidden Microphones: At the Crossroads of Journalism, Ethics, and the Law. <http://www.rtda.org/resources/hiddencamera/allstates.html>
12. Hindus, D. and Schmandt, C. Ubiquitous Audio: Capturing Spontaneous Collaboration. *Proc. CSCW '92*. 210–217 (1992)
13. Johnson, C. Privacy Lost: The Supreme Court's Failure to Secure Privacy in That Which is Most Private – Personal Diaries. *33 McGeorge L. Rev.* 129. (2001).
14. Lamming, M., Brown, P., Carter, P., Eldridge, M., Flynn, M., Louie, P., Robinson, and P., Sellen, A. The Design of a Human Memory Prosthesis. *The Computer Journal*. **37**(3), 153–63 (1994).
15. LifeLog, <http://www.darpa.mil/ipto/programs/lifelog/>
16. Lin, W. and Hauptmann, A. A Wearable Digital Library of Personal Conversations. *JCDL 2002*: 277–278 (2002).
17. Lucene, <http://jakarta.apache.org/lucene/>
18. Monty, M.L. *Issues for Supporting Notetaking and Note Using in the Computer Environment*, Ph.D. thesis, Department of Psychology, University of California, San Diego, CA (1990).
19. 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 multimedia-meeting records. *Proc. of CHI '97*. (1997).
20. Morgan, N., Baron, D., Edwards, J., Ellis, D., Gelbart D., Janin, A., Pfau, T., Shriberg, E., and Stolcke, A. The Meeting Project at ICSI. *Human Language Technologies Conference*. (2001).
21. Orwant, J. Heterogenous learning in the doppelgänger user modeling system. *User Modeling and User-Adapted Interaction*, **4**(2), 107–130, (1995).
22. Rhodes, B. *Just-In-Time Information Retrieval*. Ph.D. Dissertation, MIT Media Lab (May 2000).
23. Rogers, M.G., Bodily Intrusion in Search of Evidence: A Study in Fourth Amendment Decisionmaking. *62 Ind. L.J.* 1181. (1987).
24. Schacter, D.L. The Seven Sins of Memory: Insights from Psychology and Cognitive Neuroscience. *American Psychologist*. **54**(3), 182–203 (1999).
25. Schmandt, C. The Intelligent Ear: An Interface to Digital Audio. *Proc. IEEE International on Cybernetics and Society*, IEEE, Atlanta, GA (1981).
26. Smeaton, A.F., Over, P. The TREC-2002 Video Track Report. *Proc. TREC 11*. (2002).
27. Stark, L., Whittaker, S., and Hirschberg, J. ASR satisficing: the effects of ASR accuracy on speech retrieval. *Proc. ICSLP*. (2000).
28. Vemuri, S., DeCamp, P., Bender, W., and Schmandt, C. Improving Speech Playback Using Time-Compression and Speech Recognition. To appear *Proc. CHI 2004*. (2004)
29. ViaVoice, <http://www-3.ibm.com/software/speech/>
30. Wagenaar, W.A. My Memory: A study of Autobiographical Memory over Six Years. In *Cognitive Psychology*. **18**, 225–52 (1986).
31. Wechsler, M., Munteanu, E., Schäuble, P.: New Techniques for Open-Vocabulary Spoken Document Retrieval. *Proc. SIGIR 1998*. 20–27. (1998).
32. Whittaker, S., Hirschberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., Stead, L., Zamchick G., & Rosenberg, A. SCANMail: a voicemail interface that makes speech browsable, readable and searchable. *Proc. CHI 2002*, 275–82 (2002).
33. Wilding, E.L., Rugg, M.D. An event-related potential study of memory for words spoken aloud or heard. *Neuropsychologia*. **35**(9), 1185–95 (1997).
34. Witbrock, M., <http://infontics.com/searchengines/boston1999/witbrock/index.htm>, Lycos (1999).
35. 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).