Markup as You Talk: Establishing Effective Memory Cues While Still Contributing to a Meeting

Vaiva Kalnikaitė
Interactables
Cambridge, UK
vaiva@interactables.com

Patrick Ehlen
AT&T
San Francisco, CA
patrick.ehlen@att.com

Steve Whittaker
University of California
Santa Cruz, CA
swhittak@ucsc.edu

ABSTRACT
Meeting participants can experience cognitive overload when they need both to verbally contribute to ongoing discussion while simultaneously creating notes to promote later recall of decisions made during the meeting. We designed two novel cuing tools to reduce the cognitive load associated with note-taking, thus improving verbal contributions in meetings. The tools combine real-time automatic speech recognition (ASR) with lightweight annotation to transform note-taking into a low overhead markup process. To create lightweight notes, users do not generate the notes' content themselves. Instead they simply highlight important phrases in a real-time ASR transcript (Highlighter tool), or press a button to indicate when they heard something important (Hotspots tool). We evaluated these markup tools against a traditional pen-and-paper baseline with 26 users. Hotspots was highly successful: compared with handwritten notes, it increased participants' conversational contributions and reduced their perception of overload in the meeting, while improving recall of the meeting two months later. Highlighter also improved recall without compromising conversational contributions, although users found it more demanding.

Author Keywords
Meetings, cognitive load, speech recognition, automated transcripts, summarization, note-taking, markup, memory cuing, recall.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Experimentation; Human Factors.

INTRODUCTION
Meetings are a critical organizational practice and a key way that organizations generate knowledge and make decisions. Despite their prevalence, prior research has identified many meeting inefficiencies [23]. The result is not only frustration for participants, but reduced productivity for the organization.

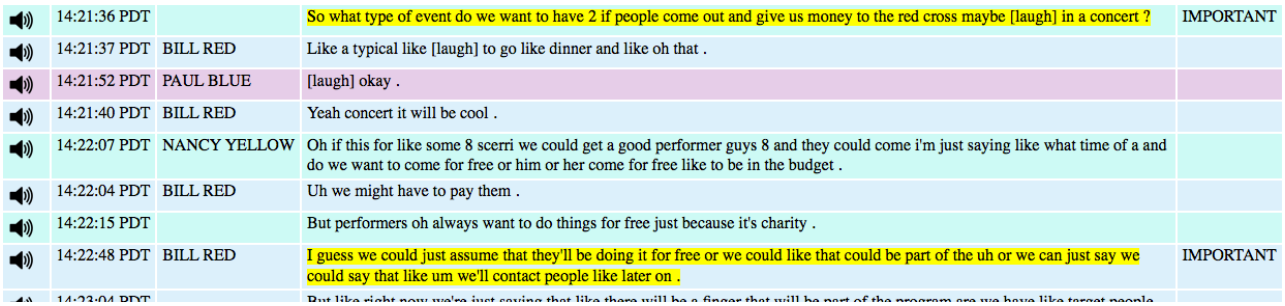
Considering these well-documented problems, meetings have proved remarkably resistant to technological intervention. Over two decades of work have seen experiments with large displays to share notes about emerging discussions [1, 4, 25, 28], recording tools to capture key points [41, 46, 47], ubiquitous methods to support idea capture [9, 35], as well as more recent forays into automatic multimedia analysis [17, 33]. In general, however, these technologies are not widely deployed.

Here we examine one simple, but important, aspect of meeting behavior: note-taking. Much prior theoretical and empirical work indicates that meetings induce *cognitive overload*. They require participants to make constant trade-offs between two competing tasks: contributing to current discussion, and preparation for later recall [27, 30]. A participant's main focus should be to actively contribute to the meeting itself. On the other hand, they also need to record important information to aid future recall. Prior work shows they are dissatisfied with their attempts to do both [20, 26, 27, 28, 42].

We designed two novel markup tools to address overload. We wanted to promote post-meeting recall without diminishing a participant's ability to contribute effectively to the meeting. To increase meeting contribution, we wanted to minimize the cognitive overhead imposed by our note-taking tools. We aimed to do this by simplifying note-taking from a demanding process of selecting and recording important information to a lightweight markup process.

Many other projects have used automatic speech recognition (ASR) to create rich transcripts of meetings that can be browsed and searched after the event [13, 14, 21, 24, 33]. This has obvious intuitive benefits: written records are created automatically, thus reducing in-meeting cognitive load and allowing participants to contribute to the meeting without having to think about taking notes. Nevertheless, these approaches have not generally been successful [43]. This may be because complete recordings omit an important characteristic of notes. Effective notes are a *personal view of a critical subset* of the meeting. People want notes to be a short summary of critical personal points, rather than a complete record of all that was said [26, 42, 46, 47].

CSCW'12, February 11–15, 2012, Seattle, Washington, USA.



	14:21:36 PDT		So what type of event do we want to have 2 if people come out and give us money to the red cross maybe [laugh] in a concert ?	IMPORTANT
	14:21:37 PDT	BILL RED	Like a typical like [laugh] to go like dinner and like oh that .	
	14:21:52 PDT	PAUL BLUE	[laugh] okay .	
	14:21:40 PDT	BILL RED	Yeah concert it will be cool .	
	14:22:07 PDT	NANCY YELLOW	Oh if this for like some 8 scerri we could get a good performer guys 8 and they could come i'm just saying like what time of a and do we want to come for free or him or her come for free like to be in the budget .	
	14:22:04 PDT	BILL RED	Uh we might have to pay them .	
	14:22:15 PDT		But performers oh always want to do things for free just because it's charity .	
	14:22:48 PDT	BILL RED	I guess we could just assume that they'll be doing it for free or we could like that could be part of the uh or we can just say we could say that like um we'll contact people like later on .	IMPORTANT

Fig 1. Highlighter post-meeting rich transcript. The transcript is generated using ASR. User selections are highlighted in yellow in the transcript, and marked as 'important'. Clicking on an utterance replays the underlying speech.

However very recent technical developments in real-time ASR [33, 39] have made it possible to develop new classes of lightweight note-taking applications. These can reduce cognitive load while allowing participants to note just those parts of the meeting that are personally relevant. Instead of 'recording everything,' these advances in ASR mean that participants can view an ASR transcript of the meeting *in real-time*, allowing them to select only parts that are of interest. Such ASR tools can therefore be lightweight. In contrast with manual note taking, users do not have to laboriously record specific content, like who said what to whom. Instead, they identify just those parts of the transcript that they think are critical for future recall, creating a personal summary of what was said.

Our first meeting markup tool, *Highlighter*, combines real-time ASR transcription with a pen-based highlighting tool. It allows users to view and mark up important parts of the transcript as it unfolds. It combines the ability to select personally important elements with the low cognitive overhead of automatic transcription.

However, it could be argued that Highlighter is still too cognitively demanding, as users have to follow the ASR transcript while participating in the meeting. Our second markup tool, *Hotspots,* simplifies note-taking even further. Like Highlighter, Hotspots allows selection of important elements of an ASR transcript. But with Hotspots, users do not view a potentially distracting transcript during the meeting. Instead they press a large button on the screen of an Internet tablet when something occurs that they want to make a note about. Pressing the button marks up points in the meeting transcript that can be accessed later.

Both Highlighter and Hotspots also provide UIs for post meeting review. In both cases, users view an ASR-generated transcript that is marked up with their personal selections and temporally indexed to the originating speech. The transcript markup differs for each tool. Highlighter depicts user-highlighted regions akin to using a highlighter pen on paper—showing user-defined highlighted regions marked up in yellow. To facilitate retrieval, clicking a highlight replays the speech indexed to the highlighted region (see Fig 1). With Hotspots, the user markup is indicated by a large red button beside the transcript text (see Fig 2). Clicking on the button replays the speech for a fixed region of one minute around the point when the index was created.

We evaluate these two new tools, comparing them with traditional pen-and-paper notes as baseline. We measure in-meeting effects on meeting *contribution* and post-meeting effects on *memory*, determining whether our tools do indeed reduce perceived cognitive load according to standard metrics. More specifically, we address the following questions concerning the benefits of ASR-based markup methods.

- *In-Meeting Contributions*: Do Highlighter and Hotspots allow people to *contribute more* to the meeting than when they use pen and paper?
- *Perceived Cognitive Overhead:* Do Highlighter and Hotspots reduce people's *perception* of cognitive overload compared with Manual Notes?
- *Post-Meeting Recall*: Are people able also to *remember* more about meeting content using the rich annotated records of Highlighter and Hotspots tools compared with pen and paper or unaided memory? And since these are novel technologies and ASR contains errors, we wanted to know how *confident* people were in the recall responses that they had generated.

There are three main contributions of this work. First, as far as we are aware, our tools are novel: ASR has not previously been deployed to support real-time markup to create personal notes. Second, our designs are motivated by prior theoretical and empirical work documenting problems of cognitive load in meetings that have not been tackled in this way before. Third, we measure in-meeting contributions and cognitive load as well as subsequent recall. Previous work has tended to focus on post-meeting recall rather than examining in-meeting effects.

RELATED WORK

Meeting systems have been the focus of much prior research. One important class of system focuses on support for personal note-taking that is used to access speech recordings. People take digital notes that are co-indexed to recorded speech to control playback. Filochat [41] combined an audio speech recorder with a PC tablet for taking handwritten notes to construct a meeting record. It

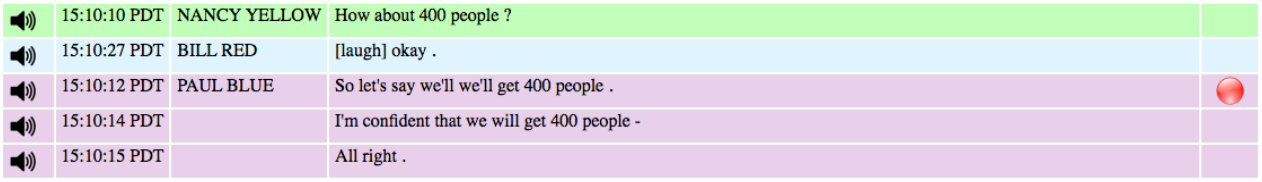


Fig 2. Hotspots post meeting UI: Transcript overlaid with hotspots showing a user selection of an important region. Clicking on the button plays back one minute of speech centered around the time when the button was pressed.

proved successful both in field and lab experiments, increasing the quality and timeliness of minutes produced after meetings. Participants were also better able to recall meeting content. Similar systems include Forget-me-not [13], Audio notebook [34] and PARC's salvaging application [26] that was successfully evaluated with patent lawyers. NotePals [11] extends this approach to allow multiple users to share notes. Dynomite [46] is a further extension: users can classify their notes into different types (e.g., 'todo', phone number, name, date, URL, etc.), allowing them to create different views onto their notes (e.g., all notes regarding 'todos' for the last month).

Other systems capture rich multimedia records of meetings. These records are subjected to visual or linguistic analysis, with browsers and search for access. The UI focus is on collaborative artifacts such as whiteboard notes, ASR transcripts or agenda items, rather than personal notes. Cutler et al.'s [9] meeting browser presents captured whiteboard images. The interface also contains a participant and whiteboard index, allowing users to jump to particular segments of the meeting. Two video components support a panoramic view and a close-up view of the current speaker. The browser also allows users to speed up playback, and to skip the contributions of selected participants. Lee et al.'s [24] system does not require a dedicated meeting room; instead, capture involves panoramic and speaker-based videos with microphones to record audio. A real-time interface allows meeting participants to examine audio and video during the meeting, as well as make notes. The system also provides an automatically produced ASR transcript of the meeting, and a set of automatically generated keyframes which can be used to navigate the meeting. Ranjan et al. [32], extend this approach using methods from television production to automatically control video views. The Ferret [40] and Jabber-2 [21] browsers also use the ASR transcript as the main UI focus along with video and speaker indices. Bett et. al. [3] also add textual meeting summaries to the transcript. TeamSpace [14] supports recording and reviewing meetings, organized around slides and an agenda.

Other work has explored browsers that allow users to access speech records of meetings in mobile contexts. Tucker and Whittaker [36] developed various ways of compressing meetings speech, including summarization, speed-up, and hybrid approaches that for example presented unimportant words much faster. Catchup [37] is a further application of this approach, presenting meeting latecomers with a spoken summary of the material they have missed.

Another approach is to augment the meeting browser with information extracted using machine learning and NLP tools. Ehlen et al. [12, 31] automatically extracted key semantic content such as action items and discussion topics from ASR transcripts of meeting discussions. User post-meeting actions on this content (e.g., correcting an action item text or clicking to add it to a to-do list) then served as implicit feedback to help retrain semantic extraction models.

Finally there is extensive work looking at note-taking in education. While note-taking is well documented to be a difficult process, several studies [20, 22, 30] show that it helps to focus attention, improving memory even when notes are not used at recall. Other studies show that the value of handwritten notes decays rapidly, aiding recall no better than unaided memory after one month [19].

NOTE-TAKING TOOLS

While this study uses *post-meeting browsers* to test recall, a key focus was on the *in-meeting* tools people used to initially record or index ideas, and whether these reduce cognitive load. We now describe the underlying tools and technologies in more depth.

CALO Meeting Assistant

The note-taking technologies used in this study were developed as part of the DARPA "Cognitive Assistant that Learns and Organizes—Meeting Assistant" (CALO-MA) program. The purpose of CALO-MA is to develop a personal assistant for meeting capture, transcription and annotation, based on real-time ASR of ordinary conversational speech that occurs during meetings [39]. Meeting participants each wear a headset connected to a Java VoIP client that transmits Speex-encoded audio to a central recognition server. Real-time speech recognition is achieved by acoustic and language models developed jointly by SRI and ICSI using SRI's Decipher recognizer, with a model trained specifically for recognition of open-domain, human-human meeting dialogue. This system produces a word error rate of 28.6%, reduced to 27.1% after unsupervised within-meeting-sequence adaptation for the CALO-MA project [39]. ASR results can be accessed via an XML-RPC protocol that provides transcriptions at a rate

close to real-time, so results are returned across the network from 1-5 seconds after speech occurs, depending on the length of the utterance to be processed. Participants use these real-time ASR results in different ways, depending on the note-taking tool being used.

Highlighter

The in-meeting Highlighter tool (Fig. 1) runs as a browser-based application on an HP tablet PC running Windows Vista. Highlighter allows participants to manually annotate regions of importance in an automatically-generated ASR transcript. Text is dynamically displayed, time stamped, and color-coded for each participant. Participants use the tablet PC's pen to highlight noteworthy regions of the transcript during the meeting (Fig 3a), in exactly the way that someone might use a highlighter pen to identify critical parts of a paper text. For post meeting review, a web-based version of the system presents the ASR transcripts along with user markup highlights (see Fig 1). Selecting a markup highlight plays back the speech associated with the highlighted region.

Hotspots

The in-meeting *Hotspots* tool runs on a Nokia N810 Internet tablet. Unlike Highlighter, Hotspots does not show the real-time transcript during markup. Instead, the tablet displays a software button that can be pressed at any time during the meeting to create a temporal bookmark to indicate when something important was said. When the button is pressed, the application creates an index associated with a time-stamp. As with Highlighter, this is co-indexed to the ASR-generated transcript. For the post-meeting review tool, the hotspot markup appears alongside the ASR-generated utterance that was spoken when the button was pressed. Selecting the hotspot causes speech playback of a one-minute region around the point specified (see Fig 2).

Manual Notes

Our experimental control for Highlighter and Hotspots was "Manual Notes," a basic pen-and-paper-based note-taking method that uses a digital pen. We used a Nokia SU-27W Digital Pen with a 5"x7" Anoto dot paper notepad. Aside from being slightly thicker than an ordinary pen, these digital pens use real ink on real paper, and look and feel the same as an ordinary pad and paper (Fig 3b). Thus the user experience is effectively identical to traditional manual notes. However, the digital pen was tooled to stream pen stroke information over Bluetooth as it occurs, recording time-stamped handwriting stroke information for every note taken, to index notes against audio. Again, a web-based version of the system was used for post meeting review. In addition to reviewing regular manual notes, participants could exploit the time indexing feature by clicking on a note to hear exactly what was said when that note was taken. Co-indexing of notes and speech has been widely used elsewhere [1, 10, 13, 19, 26, 34, 46].

While we obviously expected participants to be familiar with Manual Notes, it reproduces the cognitive overhead of traditional notes. Unlike the ASR-based tools, users have to create and organize content relating to the important aspects of what is being said, potentially interfering with their ability to contribute to the meeting.

EXPERIMENTAL SET-UP

This study was divided into two different phases, a *meeting* phase and a post-meeting *recall* phase. During the meeting phase, participants held a series of three consecutive face-to-face meetings lasting roughly three hours in total. The recall phase involved each participant answering questionnaires on a website two months later. We discuss each phase in more detail below.

Method

Participants

A total of 26 participants aged 19-60 took part. There were 7 females and 19 males. All participants took part in the meeting phase, and 19 completed the recall phase. Many previous studies of meeting technologies have used students who may not be representative of target users, as they have limited work and meetings experience. To avoid this, participants were recruited via Craigslist, flyers, and word-of-mouth to find people with more extensive and varied work experience. Recruits were all native English speakers from varied work backgrounds, including nurses, teachers, charity workers, IT professionals, a taxi driver, and a managing director. All were employed and all had experience of everyday office work and meetings. They were randomly assigned to the different participant teams and received $30 for participation in the *meeting* phase and $10 for participation in the *recall* phase.

In total, there were 9 meetings with 3 participants in each meeting. One of these meetings had one no-show participant, forcing us to co-opt a local researcher. The data collected from the researcher were excluded from the results.

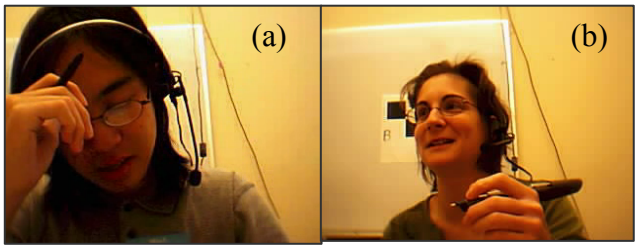


Fig 3. Study setup using: (a) the Highlighter tool to create marked up ASR transcripts; and (b) Nokia Digital Pen to create manual notes.

Meeting Phase

During the meeting phase, groups of 3 people came to our laboratory. For each group, we first described the experiment, and then ran brief hands-on tutorials to explain the note-taking technologies (Highlighter, Hotspots and Manual Notes) that participants would use during the

meeting, allowing participants time to get familiar with them. The Highlighter tool in particular, with its real-time transcript, tended to provoke curiosity and a desire to play with the tool, so we allowed participants to thoroughly explore the system before the formal experiment started. After the tutorial, participants were asked if they felt confident with using the technologies, and all reported they did.

During each meeting, three participants sat around a medium-sized round table in a meeting room instrumented for meeting capture using the CALO-MA system. Each participant wore a boom-mic headset connected to an instance of the CALO VoIP client. Audio was streamed to a central server for subsequent speech analysis. The note-taking tools were placed in front of participants on the table, and could be moved about freely.

Participants were each given nametags with pseudonym names such as "Mr. Green," "Ms. Red," etc., and were asked to refer to each other by these names throughout the experiment in order to preserve anonymity. These names were associated with each media channel in the CALO-MA system, so the transcript would attribute utterances to the correct pseudonym.

Participants were told that they were a committee charged with the goal of planning a charity event. Participants were asked to make decisions about the charity cause, the location of the event, the date, the food and refreshments, the entertainment, the itinerary, and the budget. We deliberately left the details of these items open-ended to foster active participation, discussion, and decision-making. We did not assign participants particular roles, as we believed that was likely to be one of the first decisions they might come to on their own. Before the meeting phase, participants were told they would later be given a short memory test about their discussions.

In real-life, a single meeting seldom satisfies an entire work objective, and repeated meetings with interstitial independent work are the rule. To emulate this, we divided each experimental session into 3 submeetings, lasting 20-30 minutes each, separated by intervals when participants worked independently on project objectives. We chose a single session broken into separate submeetings rather than conducting submeetings on different days to avoid repeated scheduling with potential participant attrition.

We recorded the entire conversation in each submeeting. Between submeetings, participants had a 10-15 minute period to complete independent research to address action points decided in the previous submeeting. Each participant was given a workstation outside the meeting room, located far enough away from one another to discourage chatting or discussion of details before they began the next submeeting. Their independent research entailed searching for possible project resources using the Internet with all participants starting on a search page.

The participants could also send e-mail to the "Committee's assistant"—an experimenter observing the proceedings from another room—who would occasionally reply with constraints such as, "The committee would like you to keep the budget under $20,000" to ensure the decision-making process remained active. As we wanted participants to be undistracted, they could not access their mobile phones during the entire meeting phase.

We used a within-subjects design to minimize effects of individual differences, so each participant used each note-taking tool (Highlighter, Hotspots, Manual Notes) for one submeeting. Thus a participant using Highlighter in submeeting 1 might be given the Hotspot tool in submeeting 2, and Manual Notes in submeeting 3. The tools were also varied across participants so that in any given submeeting one participant would be using Hotspots, one using Highlighter and one using Manual Notes. They were also counterbalanced for order across participants. All participants were encouraged to take note of any information relevant to their project role. While we are aware that in some meeting contexts a dedicated scribe takes notes for an entire team, observational studies of note-taking behavior show that it is more common for all participants to take personal notes [42].

After each submeeting, participants were asked to complete the standard NASA TLX (Task Load Analysis) questionnaire [16] for the tool they used in the previous submeeting. This multi-dimensional rating process uses six 7-point subscales to measure mental demand, physical demand, temporal demand, performance, effort, and frustration, combined by weighted average to provide an overall workload score to assess workload in human-machine environments. At the end of the meeting phase, we also administered a final survey where we asked open-ended questions about the tools, specifically addressing their main advantages and disadvantages as well as what could be done to improve them.

Recall Phase

A note-taking tool is only effective if it helps people to recall or reconstruct important information at a future time. So an important aspect of this study was to test participants' post-meeting recall of details, using the notes provided by each tool. We compared this to Unaided Memory; i.e., without any notes.

To give ample time for forgetting to occur, based on prior work [19] we situated the recall phase 2 months after the meeting phase. Participant recall was tested by asking each person a total of 8 questions. For six of these questions people had access to a note-taking tool and the remaining two were answered from memory. Each participant was asked two factual questions about each submeeting they participated in. To help recall, they were provided with the tool they had used in the initial submeeting. Each person had used a different note-taking tool in each of the 3 submeetings, which meant that we were able to compare

their recall using the post-meeting recall version of each tool. So for example, if a participant had taken notes with Highlighter in submeeting 1, when we tested their recall of submeeting 1, they were given the post-meeting recall version of Highlighter. We also included a final two Unaided Memory questions where we asked participants to remember meeting details without a tool. The order of the 6 tool-based questions and 2 Unaided Memory questions was counterbalanced across users. All tools were fully interactive, allowing participants to use notes to access underlying speech.

To generate the recall questions, three people listened to each of the submeetings and created factual questions based on the important decisions made during each submeeting. Example questions were: "What cause did you choose for your fundraiser?"; "What kind of entertainment did you decide on for this fundraiser?"; and, "How much money did you estimate you might make?" For each question we also generated model answers that were used to assess recall quality. We recorded binary accuracy (correct/incorrect decisions) for each question based on whether each participant's answer matched the model answers. On the rare occasions when the main judge felt uncertain about how to score a participant response it was discussed with another judge to reach consensus.

Prior work [7, 44] has shown that errorful ASR can undermine people's confidence in transcript-based tools. We therefore examined whether people were less confident in their ability to recall using our ASR tools compared with traditional Manual Notes or their own memory. After people had answered each recall question we asked them how confident they were about their answer. Ratings were generated on a 5-point scale.

RESULTS

We first report in-meeting results on spoken contributions and perceived cognitive load. Then we look at post-meeting cognitive effects on memory, and participants' confidence about their recall.

For statistical analysis, we used repeated measures ANOVAS. Before the analyses, we checked the sphericity of our data using Mauchly's test. Where the test is significant we apply Greenhouse-Geisser correction with post-hoc Bonferroni at .05.

In-Meeting Spoken Contributions

Each participant's primary goal was to contribute to the meeting. Our first question was whether the different note-taking tools affected meeting contributions. We assessed contributions by counting the number of words each person spoke in each of the submeetings. Number of words has been widely used to measure conversational contributions in the CMC literature [2, 5, 6]. It has been shown to correlate well with other contribution metrics such as turn frequency [5, 6].

Our expectation was that using Hotspots and Highlighter would reduce cognitive load, allowing participants to contribute more to the meeting. In contrast, we expected that Manual Notes would reduce contributions.

The results are shown in Fig 4. We conducted a one-way ANOVA using contribution as the dependent variable and Note-taking Tool as an independent variable. Our results show significant differences for contribution ($F_{(2,50)} = 115.2$, $p<0.0001$). From post-hoc Bonferroni tests we found that participants using Hotspots contributed more to meetings than those using Manual Notes ($p<0.04$). However, contrary to our expectations, there was no difference in contributions between people using Highlighter and those using Manual Notes ($p>0.5$). This contribution advantage of Hotspots over Manual Notes is consistent with our expectation that the single-button-press action of Hotspots reduces cognitive overhead compared with the effort required for taking Manual Notes. In contrast, Highlighter fared no better than Manual Notes, but was not worse.

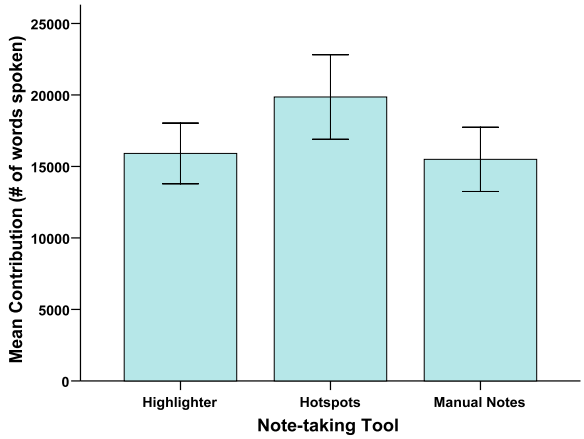


Fig 4. Mean meeting contribution for participants using each Note-taking tool. Hotspots increases contributions compared with Highlighter and Manual Notes.

In-Meeting Perceived Cognitive Effort

Hotspots increased users' contributions compared with Manual Notes and Highlighter. But is this because Hotspots reduces cognitive overhead associated with note-taking, allowing participants to participate more fully in the meeting? To answer this, we looked at the extent to which perceived cognitive overhead, as measured by the NASA TLX—matched participants' actual contribution levels. We expected that perceived effort would be reduced with our two ASR tools.

The results are shown in Fig 5. We again conducted a one-way ANOVA using the weighted average TLX score as dependent variable and Note-taking Tool as independent variable. There were significant differences among the TLX scores across all three tools used ($F_{(2,50)} = 22.5$, $p<0.0001$). As we expected, post-hoc Bonferroni tests indicated that Hotspots was perceived as less demanding than Manual Notes. However, Highlighter was perceived as *more*

cognitively demanding than Hotspots (p<0.0001) and Manual Notes (p<0.01).

To better understand the cognitive load results, we analyzed participants' comments about the different tools. Consistent with prior research on note-taking [27, 30, 42], our participants thought that taking regular manual notes was both effortful and distracting: *"It takes more time and effort to make notes. It acts as a distracting force in a meeting."*

Consistent with both the contributions and perceived effort data, people thought Hotspots demanded little effort: *"Least interruptive in meeting, least work involved."*

Some participants were positive about Highlighter, feeling it reduced note-taking effort by automatically providing content for mark up: *"[I liked it] because it translated every single word and conversation so you can mark the most helpful thing that was said."*

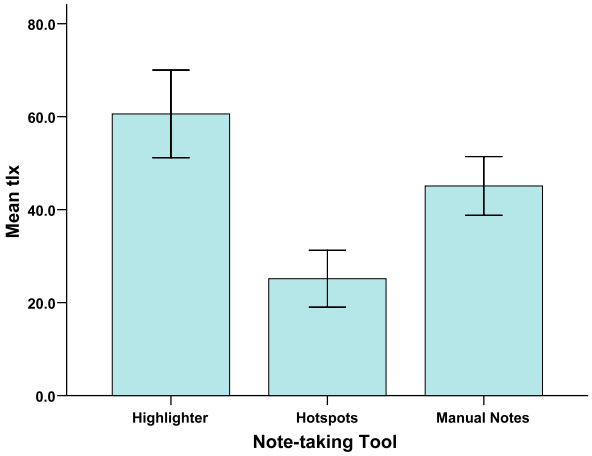


Fig 5. Perceived Cognitive Effort: Task Load Index (TLX) scores for each Note-taking tool. A higher score indicates greater cognitive effort. Perceived effort is greatest for Highlighter and least for Hotspots.

However, several other participants pointed out that Highlighter demanded additional effort: transcription lags in Highlighter increased the difficulty of identifying the relevant parts of the conversation for mark up. This problem was also exacerbated by ASR inaccuracies: *"The Highlighter took way too long to update and it was difficult finding the line of conversation that I wanted to highlight."*

These comments indicate that the advantage of Highlighter automatically generating content was partially offset by the effort needed to skim an errorful and laggy transcript to identify important regions for markup.

Overall, consistent with the contribution results, the perceived effort scores confirm that Hotspots is a cognitively lightweight annotation technique. Hotspots requires a single action to define a markup annotation, and eliminates much of the effort that goes into Manual Notes. In contrast, Highlighter, in its current form, seemed more complex. It requires users to skim the errorful transcript to identify regions of interest. Attempting to read an ASR transcript that is 1-5 seconds behind the ongoing conversation significantly increases perceived effort.

Post-Meeting Recall

So far we have seen the benefits of Hotspots in increasing meeting contributions, with Highlighter being equivalent to Manual Notes. But Highlighter and Hotspots were also intended to enhance recall compared with Manual Notes. Our next question was how well participants could *remember* details of the meeting after the event.

We examined recall of the set of factual questions generated for each meeting. Participants used the post-meeting web-based recall tools (Highlighter, Hotspots or Manual Notes) at retrieval. Each participant had initially used a different tool for capturing each submeeting, for example, using Manual Notes for one submeeting, Highlighter for the next, and Hotspots for the third. We tested recall for each submeeting, providing participants with the post-meeting version of the tool they used along with the notes they had generated during that submeeting. They therefore experienced 3 different recall tests: One for each submeeting, using the original tool they had been assigned during the meeting phase. We also test recall using Unaided Memory, with no tools provided. Our expectation was that recall would be better overall when people used the ASR markup tools (Highlighter or Hotspots) than with Manual Notes or Unaided Memory.

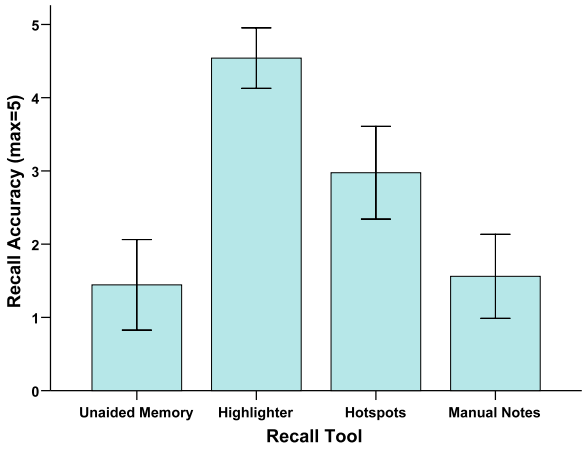


Fig 6. Mean Recall Accuracy using each Recall Tool and Unaided Memory. Recall is greater with Highlighter and Hotspots than with Unaided Memory and Manual Notes.

We conducted a one-way ANOVA using Recall Accuracy as the dependent variable and Recall Tool (Hotspots, Highlighter, Manual Notes, Unaided Memory) as the independent variable. There were significant differences in Recall Accuracy depending on Recall Tool ($F_{(3, 75)}$ = 31.9, p<0.0001), as shown in Fig 6. Post-hoc Bonferroni tests showed that Highlighter helped to generate more accurate answers than Hotspots, Manual Notes and Unaided Memory (p<0.001). Hotspots was also significantly more

accurate than both Manual Notes (p<0.002) and Unaided Memory (p<0.001). These results confirm our expectation that ASR-based tools would improve recall over manual notes and unaided memory.

Post-Meeting Confidence in Accuracy of Recall

Although people recalled significantly better using ASR-based tools compared with Manual Notes and Unaided Memory, our next question was, how confident were they about Recall Accuracy using these tools, given that ASR transcripts contained errors?

To assess people's confidence in recall accuracy, we conducted a one-way ANOVA with Confidence as the dependent variable and Recall Tool (Hotspots, Highlighter, Manual Notes) as independent variable. There was a significant difference in confidence based on recall tool ($F_{(2,50)}$ = 8.3, p<0.0001), as shown in Fig 7. Bonferroni tests showed that people were less confident about recall accuracy with Hotspots than Manual Notes (p<0.02). However, there were no significant differences between Manual Notes and Highlighter (p>0.9). The greater confidence with Highlighter compared with Hotspots may occur because Highlighter forced people to process transcripts more carefully when generating in-meeting notes, allowing people to establish the transcript was good enough to aid recall. In contrast, Hotspots required no in-meeting transcript processing, possibly leading participants to be less confident that the button-press actions would serve as effective annotations into the meeting.

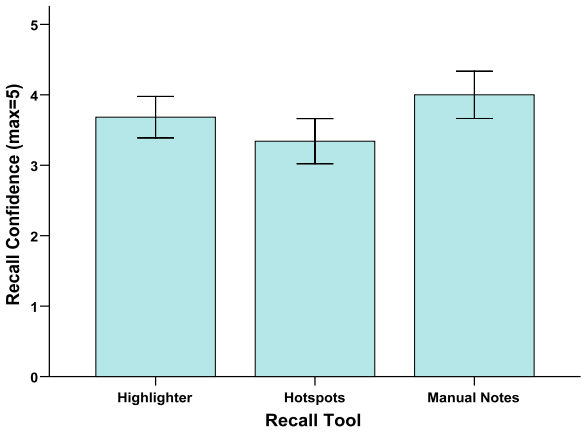


Fig 7. Mean Confidence score for each Recall Tool. People are less confident using Hotspots than Manual Notes.

Consistent with this lack of confidence, participants pointed out problems using the ASR transcript with Hotspots, noting that the transcript could be hard to interpret when ASR was poor: *"It's very difficult to read the text shown here, I can barely make anything out."* Despite this lack of confidence, recall performance was still better with Hotspots than with Manual Notes and Unaided Memory.

DISCUSSION AND CONCLUSIONS

Overall, these are highly positive results showing the clear utility of novel ASR-based markup tools in addressing the perennially tricky problem of note-taking in meetings. Prior theoretical and empirical work has repeatedly demonstrated that note-taking imposes a high cognitive load, and detracts from a participant's ability to contribute to meetings [27, 30]. However recent developments in real-time ASR allowed us to address such 'divided attention' by developing new lightweight note-taking tools. Both ASR tools support effective markup without compromising meeting contributions. For the primary task of contributing to the meeting, Hotspots increased contributions, while at the same time improving post-meeting recall compared with Manual Notes. Highlighter supported equivalent contributions as Manual Notes, and again improved post-meeting recall compared with Manual Notes.

Hotspots seemed to promote enhanced contribution by reducing in-meeting cognitive overload compared with Manual Notes (as assessed by TLX). Enhanced meeting contribution may result from Hotspots' very simple design: users simply hit a button when they hear something important. Its sheer simplicity may be crucial, because of the high cognitive demands of contributing to a meeting. In contrast, Highlighter increased perceived overload, possibly because people were focused on trying to identify regions of interest in the errorful ASR transcript, although this did not affect their contributions compared with Manual Notes.

Both Highlighter and Hotspots improved post-meeting recall compared with Manual Notes, demonstrating the benefits of the ASR transcript as a retrieval aid. However these benefits need to be tempered by the fact that transcript based methods seemed less efficient than Manual Notes or Unaided Memory, as it seemed to take participants longer to locate relevant material in the ASR transcript. In addition, despite having better actual recall with Hotspots, users were less confident about their answers than when using Manual Notes. In contrast, Highlighter elevated recall, but did not depress confidence compared with Manual Notes. Highlighter may preserve confidence because people are focused on the ASR transcript during in-meeting note-taking, allowing them to see that it provides a reasonable index into the speech recording. In contrast, with Hotspots they don't view the transcript during the meeting, possibly leading them to be unsure about its utility for recall.

These results are encouraging, especially for Hotspots, showing it to be a low-overhead tool that increased both in-meeting contributions as well as post-meeting recall. But for Highlighter, we should also ask whether its cognitive load issues are due to a nascent technology that is not quite ready for adoption. However with improved ASR models of human-human dialogue, faster processing and network speeds, and more user exposure to both speech recognition

and tablet-based interaction, some of the load disadvantage may be ameliorated, or vanish altogether.

From a methodological perspective, this is the first study to directly measure in-meeting effects on contribution and cognitive load, as prior note-taking research has tended to focus only on post-meeting recall. Still, there are limitations to this work. We need to explore technological acceptability and generality with larger numbers of diverse users outside a lab context. In addition, this study lacks one potential comparison condition: simply providing participants with an ASR transcript post-meeting, with no in-meeting markup. This would have the putative benefit of low in-meeting overhead (as no annotation is required), but providing a rich recall record. While this could be an important control to run, we do not believe it would be more effective than Hotspots and Highlighter in promoting recall, as prior work has shown the importance of *active user annotation* for inducing recall. Users who actively take notes subsequently recall material better than those who have access to verbatim meeting recordings but who did not generate notes [20, 22].

Our work also has design implications. Several user comments concerned the benefits of capturing meeting content in their own words. This suggests a modified system that is a hybrid between traditional manual notes and our new markup techniques. A simple addition to our system would be to allow users to supplement the ASR transcript in real-time with their own short comments. These personal comments could be time-aligned as in prior systems [19, 26, 41, 46]. However we do not advocate full transcript editing as in [7] because this would dramatically increase cognitive overhead, and also because our data shows that imperfect ASR indexed transcripts are still an effective retrieval aid. A final modification might be to use machine learning techniques such as entity extraction to add actions or decisions to the evolving transcript [12, 31].

Finally, there may be extensions of this approach to other applications. There has been recent interest in collecting audio lifelogs, such as sound recordings of aspects of one's everyday life [11, 29]. There are clearly problems in retrieving such records, but techniques such as Hotspots may provide simple ways to access this data, by using a simple button to mark a significant event when it occurs. It might also be possible to extend our 'divided attention' theoretical approach to lifelogging, as lifeloggers persistently have to decide whether to remain 'in the moment' or distract themselves by annotating their lifelogs to promote future retrieval [18].

In conclusion, we demonstrated the utility of two novel ASR-based markup tools. These address well-documented problems in supporting both in-meeting contributions and subsequent recall. We demonstrated that very simple methods of annotation for ASR transcripts—particularly Hotspots—can promote increased recall, while also increasing people's ability to contribute to meetings. In future work we plan to explore other situations where minimal annotation might promote enhanced memory without compromising people's current practices.

ACKNOWLEDGMENTS

Thanks to Stanley Peters and members of the CSLI Computational Semantics Lab who assisted with experimental design and setup, including Jonathan Kass, Sharareh Noorbaloochi, Justin Stimatze and John Niekrasz. This material is based upon work supported by DARPA CALO (FA8750-07-D-0185, Delivery Order 0004) funding. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA. The work was also partially supported by EPSRC grant EP/G010714/1, EU FP-6 30008, and IST FP6- 033812.

REFERENCES

1. Abowd, G.D., "Classroom 2000: An experiment with the instrumentation of a living educational environment." IMB Systems Journal, 1999, 38 (4), 508-530.
2. Anderson, A., et al. Video data and video links in mediated communication: what do users value?. Int. J. Hum Comput. Stud. 2000, 52, 1, 165-187.
3. Bett, M., et al. "Multimodal meeting tracker." In Proc. of RIAO, 2000, 324-326.
4. Brotherton, J.A., et al. "Automated Capture, Integration, and Visualization of Multiple Media Streams" In Proc. of the IEEE IMCS 1998, 54.
5. Carletta, J., et al. The effects of multimedia communication technology on non-collocated teams: a case study. Ergonomics, 2000, 43(8), 1237-1251.
6. Clark, H. H. Using language. Cambridge: Cambridge University Press, 1996.
7. Burke, M., et al. "SCANMail: Error Correction in Voicemail Transcripts" In Proc. of CHI 2006, 339-348.
8. CHIL – http://www.limsi.fr/tlp/chil.html - retrieved Sept 2011.
9. Cutler, R. et al., "Distributed Meetings: a Meeting Capture and Broadcasting System." In Proc. of ACM Multimedia 2002, 503-512.
10. Davis, R.C. et al., "NotePals: Lightweight Note sharing by the Group, for the Group." In proc. CHI 1999, 338-345.
11. Dib, L., Petrelli, D., and Whittaker, S. "Sonic Souvenirs: Exploring the Paradoxes of Recorded Sound for Family Remembering." In Proc. of CSCW 2010, 391-400.
12. Ehlen, P., et al."Meeting Adjourned: Off-line Learning Interfaces for Automatic Meeting Understanding." In Proc. of IUI 2008, 276-284.

13. Lamming, M., and Flynn, M. “Forget-me-not Intimate Computing in Support of Human Memory.” In Proc. of FRIEND 21, 1994, 125-128.
14. Geyer, W., et al. “A Team Collaboration Space Supporting Capture and Access of Virtual Meetings.” In Proc. of GROUP 2001, 188-196.
15. Geyer, W., et al. “Making Multimedia Meeting Records More Meaningful”, In Proc. ICME 2003, 2, 669-672.
16. Hart, S.G. and Staveland, L.E. “Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research.” In Human Mental Workload, 1988, 139-183.
17. Janin, A., et al., “The ICSI Meeting Project: Resources and Research.” In Proc. ICASSP 2004, 537-541.
18. Kalnikaitė, V. and Whittaker, S. “Beyond being there? Evaluating Augmented Digital Records” In IJHCS 2010, 68 (10), 627-640.
19. Kalnikaitė, V. and Whittaker, S. “Software or Wetware? Discovering when and why people use digital prosthetic memory.” In Proc. CHI 2007, 71-80.
20. Kalnikaitė, V. and Whittaker, S. “Cuing Digital Memory: How and Why Do Digital Notes Help Us Remember?” In Proc. British-HCI 2008, 153-161.
21. Kazman, R., et al. “Four Paradigms for Indexing Video Conferences.” In Proc. of IEEE MultiMedia 1996, 63-73.
22. Kidd, A. The marks are on the knowledge worker. In Proc. CHI 1994, 212-220.
23. Kraut, R. “Applying social psychological theory to the problems of group work.” In J. Carroll (Ed.) HCI Models, Theories and Frameworks, 2003, 325-356.
24. Lee, D., et al. “Portable Meeting Recorder.” In Proc. of ACM Multimedia 2002, 493-502.
25. Mantei, M. “Capturing the capture concept: a case study in the design of computer-supported meeting environments.” In Proc. of CSCW 1988, 257-270.
26. Moran, T.P. et al “I’ll Get That Off the Audio: A Case Study of Salvaging Multimedia Meeting Records” In Proc. CHI 1997, 202-209.
27. Olive, T. “Working Memory in Writing: Empirical Evidence from the Dual-Task Techniques.” In European Psychologist, 9 (1), 2004, 32-42.
28. Olson, J. S., et al. “How a group-editor changes the character of a design meeting as well as its outcome.” In Proc. of CSCW 1992, 91-98.
29. Petrelli, D., Villar, N., Kalnikaitė, V., Dib, L. and Whittaker, S. “FM Radio: Family Interplay with Sonic Mementos.” In Proc. of CHI 2010, 2371-2380.
30. Piolat, A., et al. “Cognitive effort in note taking. Applied Cognitive Psychology”, 2004, 1-22.
31. Purver, M., Niekrasz, J. and Ehlen, P. “Automatic Annotation of Dialogue Structure from Simple User Interaction.” In Proc. of MLMI, 2007, 48-59.
32. Ranjan, A., et al. “Improving Meeting Capture by Applying Television Production Principles with Audio and Motion Detection.” In Proc. of CHI 2008, 227-236.
33. Renals, S., et al. “Recognition and interpretation of meetings: The AMI and AMIDA projects.” In Proc. IEEE ASRU 2007, 238-247.
34. Stifelman, L., et al. “The Audio Notebook: Paper and Pen Interaction with Structured Speech.” In Proc. of CHI 2001, 182-189.
35. Streitz , N.A., et al. “Roomware for Cooperative Buildings: Integrated Design of Architectural Spaces and Information Spaces.” In Proc. of CoBuild 1998, 4-21.
36. Tucker, S. and Whittaker, S., “Time is of an Essence: an Evaluation of Temporal Compression Algorithms.” In Proc. CHI 2006, 329-338.
37. Tucker, S., Bergman, O., Ramamoorthy, A., and Whittaker, S. “Catchup: a Useful Application of Time-travel in meetings.” In Proc. CSCW 2010, 99-102.
38. Tur, G., & Stolcke, A. “Unsupervised Language Model Adaptation for Meeting Recognition.” In Proc. IEEE ICASSP 2007, . 173-176.
39. Voss, L., Ehlen, P., et al. “The CALO Meeting Assistant.” In Proc. NAACL-HLT 2007, 17-18.
40. Wellner P, et al. “Browsing recorded meetings with Ferret.” In: Proc. of MLMI 2004, 12-21.
41. Whittaker, S., Hyland P., and Wiley, M. “Filochat: Handwritten notes provide access to recorded conversations.” In Proc. CHI 1994, 271-277.
42. Whittaker, S., Laban, R., and Tucker, S. (2005). Analysing Meeting records: An Ethnographic Study and Technical Implications. In Lecture Notes in Computer Science 3869, Machine Learning for Multimodal Interaction, Springer, New York.
43. Whittaker, S., et al. “Design and Evaluation of Systems to Support Interaction Capture and Retrieval.” In PUC 2008, 197-221.
44. Whittaker, S., et al. “SCANMail: A Voicemail Interface That Makes Speech Browsable, Readable and Searchable.” In Proc. of CHI 2002, 275-282.
45. Whittaker, S. and Amento, B. “Semantic Speech Editing.” In Proc. of CHI 2004, 527-534.
46. Wilcox, L.D., et al. “Dynomite: A Dynamically Organized Ink and Audio Notebook.” In Proc. of CHI 1997, 186-193.
47. Yu, S., and Selker, T. “Who Said What When? Capturing the Important Moments of a Meeting”. In Proc. of CHI 2010, pp 3283-3288.