

Design and Evaluation of Systems to Support Interaction Capture and Retrieval

Steve Whittaker, Simon Tucker, Kumutha Swampillai and Rachel Laban

University of Sheffield
Department of Information Studies
University of Sheffield
211 Portobello St
Sheffield, S1 4DP, UK
phone: +44 114 222 6340
fax: +44 114 278 0300
email: s.whittaker@shef.ac.uk
(corresponding author)

Presumably man's spirit should be elevated if he can better review his shady past and analyze more completely and objectively his present problems. He has built a civilization so complex that he needs to mechanize his record more fully if he is to push his experiment to its logical conclusion and not merely become bogged down part way there by overtaxing his limited memory. His excursion may be more enjoyable if he can reacquire the privilege of forgetting the manifold things he does not need to have immediately at hand, with some assurance that he can find them again if they prove important. (Vannevar Bush).

Abstract

Although many recent systems have been built to support Information Capture and Retrieval (ICR), these have not generally been successful. This paper presents studies that evaluate two different hypotheses for this failure, firstly that systems fail to address user needs and secondly that they provide only rudimentary support for ICR. Having first presented a taxonomy of different systems built to support ICR, we then describe a study that attempts to identify user needs for ICR. On the basis of that study we carried out two user-oriented evaluations. In the first we carried out a task-based evaluation of a state-of-the-art ICR system, finding that it failed to provide users with abstract ways to view meetings data, and did not present users with information categories that they considered to be important. In a second study we introduce a new method for comparative evaluation of different techniques for accessing meetings data. The second study showed that simple interface techniques that extracted key information from meetings were effective in allowing users to extract gist from meetings data. We conclude with a discussion of outstanding issues and future directions for ICR research.

Introduction

Collocated collaboration in meetings is a key method by which organisations create and share their knowledge, and the last 20 years have seen the development of new computational technologies to support this. Early work researched group decision support systems (Poole and DeSanctis, 1989), shared whiteboards and large displays to promote richer forms of communication and collaboration (Brotherton et al., 1998, Mantei, 1988, Moran et al., 1998, Olson et al., 1992, Whittaker and Schwarz, 1995, 1999). There were also attempts at devising methods for evaluating these systems (Olson et al., 1992). More recent research on collocated collaboration has been inspired by ubiquitous computing (Streitz, et al., 1998, CHIL, Yu et al., 2000), where the focus has been on direct integration of collaborative computing into existing work practices and artefacts. While much of this prior work has addressed support for *real time communication* by providing richer interaction resources, another important research area is *interaction capture and retrieval* (ICR) and this is our focus here.

ICR is motivated by the observation that much valuable information exchanged in workplace interactions is never recorded, leading people to forget key decisions or repeat prior discussions. Its

aim is to provide computational techniques for analysing records of interactions, allowing straightforward access to prior decisions, actions and discussions (AMI, ICSI). Interaction capture is clearly a difficult problem. A great deal of technology has already been developed to support it (Brotherton et al.1998, Mantei, 1988, Moran et al., 1997, Moran et al., 1998, Whittaker et al., 1994b). So far however, there has been little uptake of ICR technology and the most common interaction recording techniques used in meetings today are still pens, notebooks and more recently laptops.

This paper explores two potential related reasons for this lack of uptake.

One possibility is that these technologies *fail to address user needs for ICR*, i.e. they do not capture the data that users want in a way that makes it accessible. There has not been a great deal of research into user needs for ICR. We therefore revisit the issue of user requirements reporting a study that explores user needs.

A second (less serious) possibility is that current systems *provide only rudimentary support for ICR*, requiring extensive modification if they are to truly meet user needs. To determine whether this is the case we need to evaluate ICR systems, and we present two studies using different methods for evaluating ICR technologies. The first uses *task-centric evaluation* to determine how well a state of the art system supports the user needs identified in our requirements study. The second study develops and applies a method for *comparing multiple techniques* for accessing the *gist* of a recording.

The structure of the paper is as follows. We first review previous ICR technologies, along with prior work investigating user requirements for ICR. We then report 3 studies of our own. In the first, we present an ethnographic study looking at user needs for ICR, based on an analysis of current techniques for creating interaction records. On the basis of those results, we carry out a requirements-centric evaluation of a state-of-the-art ICR browser, identifying the main weaknesses of current browsers. We then present a further study where we develop a novel technique for comparative evaluation of multiple ICR browsers. We conclude with a discussion of the implications of our research for future technologies and evaluation methods.

Previous ICR research

ICR Systems

Table 1. Taxonomy of meeting browsers and typical indexing elements used in each class

Perceptual	Semantic
<i>Audio</i>	<i>Artefacts</i>
<ul style="list-style-type: none"> • Speaker Turns • Pause Detection • Emphasis • User determined markings 	<ul style="list-style-type: none"> • Presented Slides • Agenda Items • Whiteboard Annotations • Notes - both public and private • Documents discussed
<i>Video</i>	<i>Discourse</i>
<ul style="list-style-type: none"> • Keyframes • Participant Behaviour 	<ul style="list-style-type: none"> • ASR Transcript • Named Entities • Mode of Discourse • Emotion

One way to categorise different ICR systems is in terms of the *focus* of their browser. Focus is defined to be either the main device for navigating the data, or the primary mode of presenting the meeting data to the user.

We identify 4 main classes of browsers, as shown in Table 1. There are browsers whose focus is *audio*, including both audio presentation (Degen et al., 1992, Hindus and Schmandt, 1992) and navigation via audio (Arons, 1997). Others focus largely on *video*; again including both video presentation systems (Girgensohn et al., 2001) and those where video is used for navigation (Christel et al., 1998). The third

class of browsers presents *artefacts* of the meetings. Meeting artefacts may be notes made during the meeting, slides presented, whiteboard annotations or documents examined in the meeting. All of these can be used for presentation and access. A final class of browser focuses on derived data such as a transcript generated by applying automatic speech recognition (ASR) to the interaction recording. Other derived data might include: entities extracted from the recording (such as names, dates or decisions), emotions, or speech acts. We call this final class, *discourse* browsers because their focus is on the nature and structure of the interaction. We refer to audio and video indices as *perceptual* since they focus on low-level analysis of the data using signal processing methods. Artefacts and derived indices are referred to as *semantic* since they perform a higher-level analysis of the raw data.

Perceptual and semantic systems have different underlying user models. Perceptual systems assume that users will access data by browsing audio or video media selecting regions of interest using random access. In contrast, semantic systems provide higher levels of abstraction, allowing users greater control using search, or by accessing key parts of the meeting (such as decisions and actions).

Audio browsers

Speech is difficult to browse given its sequential nature (Whittaker *et al.*, 2000, 2002, 2004). Some *audio* browsers therefore present a visualization of signal energy or a representation of different speakers, allowing users to navigate to points of interest within the recording. Others provide users with no limited or no visual feedback.

Both Kimber *et al.* (1995) and Hindus and Schmandt, (1992) described browsers whose primary means of navigation is via a visual index generated from speaker segmentation. Degen *et al.* (1992) developed an indexed audio browser designed for visually reviewing recordings made with a personal tape recorder. The tape recorder allowed users to mark salient points whilst recording. The marked recordings are then digitised for review on a computer. The computer interface affords several methods of browsing. Users can navigate using the markings they made during the recording phase. The visual representation shows amplitude against time, displayed as a vector or colour plot. Users can also zoom in and out of this display and have the ability to speed up playback.

A key element of all the above browsers is that the visual representations allow users to immediately see aspects of the structure of a meeting. This view, however, is dependent on the browsing environment allowing complex visual representations to be displayed. But there are situations which do not allow for such visual feedback (e.g. when using a phone or other device with limited display), so 'pure' audio browsing requires a substantially different interface.

SpeechSkimmer (Arons, 1977) is a system for interactive 'skimming' of recorded speech. Skimming is defined as system-controlled playback of samples of original audio. *Speechskimmer* has a four level skimming system, with each level compressing the speech further, while still retaining salient content. The first level is unprocessed playback, the second shortens pauses, whilst the third level plays back only speech which follows significant pauses (on the grounds that such pauses signal the beginning of new 'audio paragraphs'). The final level uses an emphasis detector to select salient segments of the speech to present to the listener. On top of these skimming levels, is a mechanism which allows the playback speed to be altered whilst maintaining speaker pitch. In this way the playback speed can be increased without a significant loss in comprehension. The interface also allows users to skim backwards in a recording - in this mode, short segments of speech are played forwards, but in reverse order.

Roy and Schmandt (1996) describe a portable news reader implemented in a small, Walkman style device. The interface allows listeners to playback a news report and navigate using pre-defined jump locations, again computed from an analysis of pause lengths in the audio. Users preferred simpler, more controlled interfaces, preferring manual skims via jumping, rather than having software controlled skims. The device also implements a form of speed-up similar to that described above, with users able to select from three different playback speeds.

In general most of the research into audio browsers has focused on proof of concept demonstrations, and there has been little technology evaluation. Where evaluation has been done (e.g. Arons, 1997; Roy and Schmandt, 1996), this has largely been confined to informal methods such as the use of heuristic evaluation, or feedback from small user groups.

Video Browsers

There are no examples of systems that present ‘pure’ video data without audio. However a number of systems use video as their central UI focus, and these we refer to as video browsers.

Girgensohn *et al.* (2001) describe video interfaces centred around the use of *keyframes*. Keyframes are significant static images that have been automatically selected from continuous video. They can be used to navigate between visual scenes. Keyframes are chosen according to an importance score, depending on the rarity and duration of each shot. Frames are sized according to their importance (so that keyframes of higher importance are larger) and are placed linearly on the page. The resulting interface is similar to a comic book or Japanese Manga drawings. This method can be used to produce a single summary of a full meeting and the user can playback salient portions of the meeting by selecting keyframes, or by choosing a point on a horizontal time line.

Foote *et al.* (1998) describe a simple video browser with two primary modes of navigation. Users can access index points precomputed from properties of the audio and video. The same indexing, when converted to a continuous ‘confidence’ measure can also be used to control the playback speed, based on whether a part of the meeting contains significant events. For example, points in the video containing significant gestures are first identified using gesture recognition. This analysis is then used to control playback speed, so that portions of the meeting with highly significant gestures are played at slower speeds.

A more complex video focused meeting browser is described by Lee *et al.* (2002). A benefit of this system is that it does not require a dedicated meeting room; instead, capture is performed by a single device, using a camera to capture a panoramic video of the meeting and four 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 browsing interface has a large number of navigational options. Central to the interface is the video screen, showing both the panorama and a close-up of the currently speaking participant. Users can navigate via a number of indexes, including representations of speaker transitions, visual and audio activity. 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. The interface also allows the user to review any notes made during the meeting and to examine any artefacts produced from the meeting.

Video is usually supplemented with other browsing devices and is rarely the sole means of navigation. More critically perhaps, meeting data often does not contain the salient visual events that make visual events truly useful for video browsing, as participants’ visual activity in meetings tends to be rather restricted. So, while there is potential for making use of video as a navigational tool, its value is not yet determined (Christel *et al.*, 1998).

Artefact Browsers

We use the term artefact to describe any physical item recorded during a meeting which isn’t audio or video. Artefact browsers fall into two subclasses: those which focus on *slides* presented and those which focus on *notes* taken by meeting participants. An important difference between this class of system and video or audio browsers is that artefacts are usually *searchable*.

Cutler *et al.* (2002) describe a meetings browser in which the central component is captured images from a whiteboard. The interface also contains a participant and whiteboard index, allowing users to jump to particular segments of the meeting, or to review segments which relate to specific elements of the whiteboard annotations. Furthermore, two video components are included - a panorama of all the participants and a close up view of the current speaker. In addition, the browser also allows users to speed up playback, and to skip the contributions of selected participants.

Brotherton *et al.*’s (1998) system attempts to extract useful information from classroom lectures. The interface shows the slides used during the presentation. Slide transitions are indexed, allowing the user to jump to segments of the presentation, e.g. the speech accompanying each slide. The slides can also be manually annotated both during, and after the presentation and this information is also indexed. Users can also search through transcribed audio, slide text and annotations. In this way, students are able to

see how different topics relate to each other, as well as being able to locate specific information from a series of presentations.

A related system TeamSpace (Geyer *et al.*, 2001) supports the organising, recording and reviewing of meetings. The interface consists of two main components. The first presents an overview of the full meeting and the second, a portion of the overview showing details of the region the user has selected. The second main component for browsing is a tabbed pane containing annotated slides, agenda items and video displays. The slide view has an index showing each of the meeting slides, so that users can use these as a navigational aid to jump to relevant portions of the meeting. The meeting agenda can also act as an index for the meeting, with progress through agenda items being tracked and signalled in the other interface components.

The browsers described above have focused their attention on presenting *community artefacts* - those which can be altered or viewed by *all* meeting participants. Another set of artefact browsers examine more *private artefacts*; specifically, they make use of private notes made by *specific individuals* as a means of *personal* indexing and browsing.

Whittaker *et al.*'s (1994b) *Filochat* system combines an audio recorder with a tablet for taking notes as a means of constructing a meeting record. Notes are temporally coindexed with audio, so that clicking on a note accesses the speech that occurred when the note was taken. The tablet PC provides a virtual notebook that allows users to store several pages of notes and organise them into sections. Users can then use the notes they have taken to jump to the relevant portion of the conversation. The system proved successful both in field and lab experiments, increasing the quality and timeliness of minutes produced after meetings.

A related system was Dynamite (Wilcox *et al.*, 1997) a pen-based system that allowed users to take notes associated with meetings. A critical feature of the system was that users could classify their notes into different types (e.g. 'todo', phone number, name, date, URL, etc.), which allowed users to create different views onto their notes (e.g. all my notes about 'todos' for the last month). Note-taking behaviour was also used to control audio recording so that although all audio was recorded only portions that were accompanied by note-taking activity were saved – on the grounds that these were more likely to be of importance.

The use of notes to assist recall of meetings was also investigated by Moran *et al.* (1997). They examined meetings chaired by a single person, in which audio and notes taken both on a shared whiteboard and by the chair were recorded and timestamped. The records were then used by the chair to make technical decisions on the basis of what was said at the meeting. Moran *et al.* looked at how the chair used the meeting record over a large period of time and identified not only how the meeting record was used, but also how the chair's use of the meeting record changed over time. In particular the chair developed shorthand conventions to annotate events of interest. The chair would often annotate his notes with the word "ha", to indicate something interesting had occurred and that it would be useful to revisit this section of the meeting during the review process. In general the record was used successfully to supplement the rudimentary notes the chair had taken during the meeting.

User notes combined with speech are a powerful means of indexing meetings. They serve two functions. On the one hand they are a user-defined index to audio events that the user thinks are significant. On the other hand, the notes can be cryptic as the notes also provide ready access to the accompanying speech if they later prove to be hard to understand. Unlike audio and video browsers, artefact browsers have been systematically evaluated, and personal notes in particular have been shown to the highly successful in supporting meeting recall.

Discourse browsers

A fourth class of browser focusses on *derived data forms*. Since analysis of meeting data relies on the nature and structure of conversation, this final class is concerned with browsing *discourse*. We include browsers whose focus is the automatic speech recognition (ASR) transcript (Tucker and Whittaker, 2004), and those which focus on the temporal structure of discourse between participants (Wellner *et al.*, 2004). Since they make use of both raw and derived data, browsers in this category tend to have a more complex interface than those discussed in the previous classes. But this increased complexity also allows for complex interface components, such as a search facility based on the transcript.

Bett *et al.* (2000) describe a meeting browser in which the transcript is given prominence over video components. The browser also allows the user to construct audio, video or text summaries, using text processing, for complete meetings or salient segments of the meetings. The summary is based on the transcript data and the audio and video streams are segmented to fit with the reduced transcript. The browser also supports search of a large meetings archive and indexing of discourse features and detected emotions.

The *Ferret* browser (Wellner *et al.*, 2004) features the transcript as the main UI focus, alongside video and participant indexes. A key feature of the browser is that additional temporal annotations can be added or removed at will. For example, it is possible to add automatically derived agenda and interest indices whilst browsing a meeting. As with other browsers, users can navigate through the meeting by clicking on the transcript or by using the derived indices. The index view is customisable and can be viewed at a variety of different zoom levels.

The *Jabber-2* system described by Kazman *et al.* (1997) has many similarities to the *Ferret* browser. Central to *Jabber-2* is the temporal view, showing the involvement of each participant. There is also a topic browser displaying a set of keywords, derived using text processing, showing the content of the current meeting organised by topic. The system also presents a contribution overview by plotting a graph showing the amount of involvement for participant in each segment of the meeting. A previous version of the same system *JabPro*, contained a video component and allowed users to search the meeting transcript to identify where in the meeting keywords occurred. Kazman *et al.* also conducted an evaluation of the browser where they compared participants and non-participants accessing meeting records. For non-attendees, they found little utility for more advanced system features such as the topic browser.

A different approach is taken by Lalanne *et al.* (2003). Here, the ASR transcript is supplemented with audio and video controls, as well as a view of documents currently being discussed (e.g. agenda or position papers). The meeting is also indexed according to participants' contributions as well as properties of the documents and discourse occurring throughout the meeting. Every interface component is synchronised, so that any change or transition in one component is automatically reflected in all the other components of the interface, e.g. relevant parts of the position paper, slides and agenda are all highlighted when participants begin to discuss these.

User Requirements and Problems with Existing ICR Technologies

While a great deal of research effort has been directed into the development of new ICR technologies, much less work has been done into determining user requirements or the systematic evaluation of these technologies. The remainder of the paper tackles these issues. What little requirements work that has been done has used 3 main methods: (a) large scale surveys of tools and memory processes; (b) query elicitation; (c) analysis of existing recording practices.

Jaimes *et al.* (2004) used an internet questionnaire collecting data from 519 participants to assess the current use of tools to review meetings, specifically focusing on the utility of visual information in accessing meetings. Their survey showed that the most common retrieval aids for meetings are: documents such as agendas or position papers (used by 88% at one time or another), personal notes (82%), minutes (79%), asking someone (75%), whiteboard notes (48%). Digital and analogue recording media such as photos, videos and audio recordings had never been used by 75% of their population. A second part of the survey looked at the benefits of verbatim records. These included: verification (e.g. checking what someone said), understanding (e.g. listening to missed content), reexamination (viewing in a different context), keeping records (logging discussions and events) and recall (remembering ideas that weren't explicitly part of the meeting content).

Follow-up interviews determined what specific information (e.g. meeting participants, location, topics discussed etc.) could be recalled after participating in a meeting. One important observation was about the nature of participants' retrieval processes. For meeting participants, accessing information consisted of *reconstruction* rather than pure retrieval. For example meeting participants may be able to remember who first made a point, or that the point was made in a presentation, while not being able to recall the

exact details of that point. The fact that users often have access to such partial information has important implications for design. Providing access to this partial information in the interface might allow users to recall information they may not otherwise have been able to remember. A clearer understanding of the nature and type of this partial information should allow it to be used to better cue recall.

A second method of generating user requirements was described by Lisowska (2003). She generated use cases from a literature survey of typical uses of meeting browsers. Participants were asked to select a use case and then generate questions they would ask a meeting browser system for that case. An analysis of the 300 responses revealed several properties that a meeting browser should possess but also highlighted a noticeable difference between the questions asked by project members and non-project members.

Another approach to determining user requirements is to analyse current recording practices to identify the main functions of these recordings as well as problems with the technologies that people are currently using. Whittaker *et al.* (1994b) carried out two sets of interview studies. They examined the use of tape-recorders and pen-and-paper note-taking practices, probing the effectiveness of these technologies for constructing effective records. They first interviewed 23 people who used tape recorders to create meeting records, looking at their main benefits and disadvantages. Users emphasised the benefits of a verbatim record of what was said, and the freedom that recording allows them to concentrate on participating in discussion. But users also encountered major problems trying to transcribe and access large amounts of recorded speech. Access seemed to be a laborious trial and error process, often using accompanying handwritten notes. Users either transcribed the whole recording or listened to it to try to extract important points. There was little use of technical indexing facilities such as the tape counter for access.

A second interview study of note-taking in meetings in 28 professionals found that most users saw note-taking as serving to capture and distil key elements of what was said in meetings. However, the process was error-prone and the majority of users (70%) reported problems with their current note-taking practices. The major ones they identified were: the failure to note facts that turned out to be crucial afterwards; illegible names; uninterpretable abbreviations; the fact that often there was not enough time to write notes with the consequence that vital information was missed; the fact that note-taking reduces the ability to participate in meetings; and most generally, notes can be inadequate for detailed understanding of what was said or agreed after some time has elapsed. There were large individual differences in the number, style and stated importance of notes. These differences largely seemed to depend on the interviewee's job category. Certain people, such as journalists, lawyers, marketers and students took extensive notes, and encountered more problems than the other professions.

Although these studies offer some insights, it is apparent that we lack systematic information about user requirements. We therefore carried out a study to look directly at critical user tasks associated with ICR using a combination of the above techniques.

Gathering User Requirements

Participants, Methods and Study Context

The study was mainly ethnographic using multiple methods and techniques. Its aims were to identify: (a) user tasks when creating records; (b) the types of records that users created; (c) the main functions these records serve; (d) the limitations of current records.

The setting for our fieldwork was two UK service firms, one responsible for national and international mail deliveries and the other for supplying software services. In each firm we studied a core team through a sequence of multiple meetings. We chose to follow two teams in *repeated* interactions, rather than a large set of individual meetings, as an important issue concerned how information in earlier meetings was invoked and followed up on in later meetings.

The core teams had 5 and 7 members, and the target meetings had between 3 and 16 participants. This numerical discrepancy arose because not all team members were able to attend all team meetings, and in the software company some meetings included customers from outside the organisation. In the

delivery company, meetings were held on a weekly basis. Their main objective was reporting and team co-ordination, as the team worked through issues arising during the previous week's operations. In the software company, meetings were between customers and suppliers. The main objective was to iron out difficulties associated with supply and delivery of services. Both sets of meetings were task-oriented rather than being about idea generation and they tended to be structured around written agendas. In both cases core participants were familiar with each other, having worked together for over 6 months.

We collected many different types of data, starting with observations of participants' behaviour during meetings - when they took notes, when they talked and what they noted down. We carried out interviews with participants before and after meetings, and we analysed private notes and public minutes of the meetings. Overall we observed 7 separate meetings, generating 12 hours of observation, along with 25 hours of interviews over the course of a three-month period. We also made audio recordings of a subset of 3 meetings and transcribed these recordings. We had participants identify important parts of those meetings which we then compared with their private notes and official meeting minutes, asking participants to explain whatever discrepancies arose between what they said was important and what they had actually noted down. Our principal observer had previously worked at the delivery company and had also dealt with the software company and was therefore familiar with both companies, allowing her to gain ready access to employees.

One issue that required careful negotiation was confidentiality. At the outset, we explained the goals of the study and asked all participants whether they were comfortable being observed and recorded. All transcripts were anonymized - removing all identifying information such as participant, supplier and dealer names. All participants were informed that they could stop recordings being made at any point. They were also given copies of the meeting transcripts and observations and asked whether they wanted information to be excised from the transcript and the original record.

We begin by describing users' main tasks and the nature and functions of *public records* of meetings such as minutes. We identify the limitations of minutes, and how participants respond to these problems by taking *personal* notes in meetings. We describe the nature and functions of *personal records*, as well as the problems users experience with these.

Results

Key User Tasks

We were able to elicit a set of core user tasks from observations, notes and user comments. Users were focused on extracting two main types of information from the recording:

Gist – where the goal was to reconstruct or build up a *general* picture of what happened in the meeting, including:

- the goals of the meeting, main topics and main contributors
- key points, decisions and actions
- overall atmosphere

Specific Facts – where the focus was on much more specific information, including:

- specific comments, opinions, contributions

We elaborate these activities in more detail below, in our analysis of public and private records. There were also two main situations in which users wanted access to records (a) as a meeting attendee who was later trying to reconstruct information that they partially remembered, (b) as a non-attendee who was trying to discover information about a meeting that they had missed. The main difference between the two is that in the attendee case, users have access to partial information so the process is one of reconstruction, whereas non-attendees have no prior information available to structure their access.

The Nature and Function of Public Meeting Records

Our participant observations, analysis of minutes and interviews revealed that public records such as minutes had four main functions:

- To track group progress.
- To serve as a public record of past actions and decisions.

- To remind people about their commitments.
- To resolve disputes about commitments.

Minutes are an abstract record of attendees, group decisions, past actions and future commitments. They also document whether previous commitments and actions by group members have been carried out. On some occasions, they might include background information relating to a decision or action, but this tends to be the exception rather than the rule. Minutes are also general: they document major decisions and actions, rather than focusing on specific aspects of the meeting.

Minutes were used in a variety of ways. People firstly used them as a public record against which to track group progress in order to determine whether recent actions or commitments had been carried out. Meetings often began with a run through the previous meeting's minutes. Participants would quickly review the main items minuted from prior meetings, check whether actions had been carried out, and explore whether there were follow up items resulting from those actions. In this sense minutes help individuals co-ordinate their own actions with each other and with what was publicly agreed. One manager commented:

It is like a checkpoint for me, just to make sure what we are doing is what we agreed we'd do at the last meeting. Are we still on track with what we said we'd do?

Minutes also serve as a long-term archive of the group's commitments and actions. Very occasionally teams would be asked by management or clients about past events or commitments, and here the minutes were used as the document of record, stating what had been decided or what had been done about a particular issue.

A slightly different function of minutes was to remind people about their commitments. Here minutes serve as a public 'todo' list for the various group members detailing their individual commitments. If those commitments have not been met, then other team members or the manager will invoke the minutes as way of enforcing that the action is carried out.

Minutes are an important record of what we said we'd do and when we said we'd do it. I go back to them when people aren't happy that a particular situation has occurred because someone hasn't done something they said they'd do.

This highlights a critical aspect of minutes - their use as an implicit contract between the different group members about the actions they each agreed to carry out. Managers or other team members will refer to the minutes as a way of questioning individuals about the status of one of their commitments.

It is generally where something has not happened or something has not gone to plan. Probably the key thing from my point of view is say two or three weeks after a meeting has taken place if somebody had agreed to take an action or do something and the situation has not improved or someone has escalated a problem I'll refer back to the meeting minutes to find out what was agreed and therefore what somebody should have done.

The contractual nature of the minutes is further underlined by the practice we observed of having people 'sign off' on the minutes. Participants were encouraged to review the minutes of the prior meeting to correct discrepancies between what was written in the minutes and what participants recalled being discussed. These discrepancies need to be resolved before the minutes can become the official document of record. One manager semi-jokingly used quasi-legal language when giving team members the opportunity to approve or challenge the minutes: 'are these minutes a true and fair record of what happened on 24/11?'

In a similar contractual way, minutes are used as the document of record to resolve group disputes. There were occasional disagreements between team members about what had been decided. Usually the disputes focused on who had agreed to undertake an action item. When this happened, rather than relying on memory, the minutes were used to determine what was agreed:

I think you find over time that some people are generally more honourable about what they have said than other people, so some people you can trust what they have said and that they'll do things, other people will change their mind over time about what they said they would do and I think perhaps where I have felt in the past that someone has said something in a meeting and then backed away from it or not done what they have said then I'll generally capture what they have said or done and I'll make sure it is minuted.

The Limitations of Public Records

Although minutes had clear benefits in serving as a group contract and memory aid, our participants felt they nevertheless had several critical failings, (summarised in Table 2) including:

- Not all meetings had minutes taken.

- The minutes are occasionally inaccurate.
- The minutes lack sufficient detail to allow participants to carry out personal actions or to allow non-attendees to determine what went on in the meeting.
- They are selective sometimes omitting politically sensitive information.
- They are not timely.
- They are laborious to produce.
- They don't capture the experience of being in the meeting.
- They don't capture more peripheral aspects of the meeting such as 'awareness' information that is relevant to the group's functioning but not directly related to a decision or action.

Type of Record	Functions	Main Problems
Public Record (Minutes)	Group Todos -(actions/decisions) Summary/Gist Group Archive (history)	Not timely Lacks context & completeness Requires effort to produce No minutes taken
Private Record (Personal Notes)	Personal Todos -(actions/decisions) -(context for actions) Briefing for non-attendees Personal Archive	Esoteric Detracts from ability to contribute Failure to note key facts

Table 2: The Function and Limits of Public and Private Meeting Records

Only 56% of the meetings that we observed had minutes taken. This seemed to depend on factors such as importance, meeting context and meeting type. Minutes were taken more often in the software than the delivery company. A possible reason for this was that the software meetings were contractual in nature involving discussions between customers and suppliers, where various promises were being made about what services would be delivered. Both parties felt that it was advantageous for decisions and commitments to be a matter of record. When there were no minutes, participants relied on the manager's notes if these were available, or a combination of different team members' personal notes.

However, even when minutes were taken, participants complained that these had significant limitations. They pointed out that minutes were sometimes inaccurate. All participants routinely checked meeting minutes against what they had personally noted or remembered. They stated this was because important information was occasionally misrepresented in the minutes. These inaccuracies could arise because a discussion was complex, or poorly structured, or when the official minute-taker was not an expert in the topic under discussion. Inaccuracy is clearly a serious problem if minutes are being used both as a group archive and a contract between members about what they have agreed to do.

if someone else had taken a key action in a meeting I would make a note of that possibly, mainly to compare with the minutes of the meeting when they come out to check whether the meeting minutes were accurate particularly when I think something important or significant has been agreed.

Another major problem was that public minutes often did not provide enough information to allow participants to carry out their individual commitments. Bare statements of action items and who was responsible for each, often did not provide enough contextual information, making it hard for participants to carry out their action items.

I take notes because the minutes sometimes don't tell me everything I need to know about my own actions.

Lack of detail also meant that it wasn't always clear even to attendees exactly what had happened in the meeting:

if you don't have a more detailed record of the meeting that sometimes you lose the meaning, you lose a lot of the richness about what happened so if you just see actions it doesn't always give you a clear view of what was discussed.

This lack of detail made it even harder for non-attendees to use the minutes to discover what went on at a meeting.

Well normally minutes aren't enough, you need someone to give you a briefing afterwards, because the minutes don't tell you everything that has gone on and the discussions that took place.

The minimal nature of minutes also made it hard to revisit prior decisions or reuse prior work. We asked participants whether they ever referred to past minutes when a related issue had occurred in a prior meeting. Again the minutes were felt to be too cursory - providing insufficient context about what had been discussed to make them useful.

Another limitation is that minutes can be selective, containing deliberate omissions such as when there is a politically sensitive discussion. We noted several such 'off record' discussions which were sometimes prefaced by instructing the minute-taker not to minute subsequent comments. Though these off-record comments often contained significant information (on one occasion, unofficial confirmation of a £3.6 million contract was discussed), this was not recorded and was therefore unavailable to non-attendees.

Another factor that undermined the utility of minutes as a group 'todo' list was that they were not timely, often taking several days to produce. If individuals relied on the minutes as a reminder about their outstanding actions, then several days might elapse before they can begin those actions. This not only left them with less time to execute actions before the next meeting, but also increased the likelihood that they might forget important details associated with those items, especially as minutes tended to record minimal information about each action.

A further problem with minutes is that they are laborious to produce. A meeting participant has to be delegated to take detailed notes, reducing their ability to contribute. In addition, transposing these detailed notes, possibly checking their accuracy with various stake-holders, all means additional work.

A less frequently mentioned limit of minutes was they didn't recreate the feeling of being in the meeting. Two participants mentioned wanting records that were richer than descriptions of decisions and actions, saying they wanted to be able to reconstruct the meeting context and what it felt like to be at it:

it's just not remembering a list of some key points from a meeting but being able to transport yourself back in some instances if you are discussing what happened, so it is more of a transporting your memory back into the actual situation to remember the actual discussion to remember what actually happened.

Another limitation of minutes related to their focus on decisions, actions and commitments. Participants pointed out that a key part of meetings is to provide awareness information, unrelated to specific actions or decisions but which provides a backdrop to the group's activities. Examples here included personnel changes in other groups or high level management. A related point was that an important function of meetings was to establish a culture or modus operandi for the group and that this type of information never appeared in the minutes.

The Nature and Function of Personal Meeting Records

Participants addressed some of the limitations of public records by taking their own personal notes. 63% of our informants reported that they 'always' took personal notes. The remaining 37% said that they 'sometimes' did so, and pointed to various factors such as chairing meetings - which prevented them from taking their own notes. The main reasons for taking personal notes was to record personal action items, but more importantly to provide the *context or background information* needed to allow these actions to be correctly executed. Such contextual information was usually missing from public records.

It was clear that personal records were highly valued. All informants reported referring back to meeting records for each meeting at least once - with 75% of doing accessing a set of notes 'frequently'. Another sign of their importance was that informants took great care to ensure that they were accurate. Half of them 'occasionally' rewrote their notes. Others stated a desire to rewrite their own notes but lacked the time to do this. They also took care to preserve their notes; 75% filed meeting records, keeping these records for a year on average.

Personal notes generally had a less predictable structure than minutes. Like minutes, personal notes mentioned important decisions, names, dates and actions. However one major difference was that personal notes reflected the note-taker's personal perspective, unlike the minutes - which were a general and often formulaic record of what transpired in the meeting.

My notes are a reflection of the things that interest me, the things that are of a particular interest to me in the meeting. When people are talking but I'm not interested then I don't note anything. My notes are subjective.

These comments were also supported by our observations, where it was clear that different participants took personal notes at different times and about different agenda items.

In their personal notes, participants supplemented information about group decisions and actions with detailed factual information they thought they might forget, or which was relevant to the execution of their own personal actions. Occasionally, people might note down actions associated with others if these had relevance for their own activities.

We analysed the content of people's notes. We classified each note depending on whether it concerned decisions, actions, or contextual information. Consistent with minutes, we found that a significant proportion of personal notes concerned decisions (19%) and actions (48%). However in contrast to minutes, we found that 30% of personal notes concerned comments supplying context for actions.

Another characteristic of personal notes is that they tend to be cryptic, often consisting of a few words about a topic. There are two main reasons for this. Firstly participants are aware that taking detailed notes detracts from their ability to contribute to the discussion, so they write as little as possible. Secondly personal notes are intended to be associative triggers or reminders for the note-taker, rather than verbatim transcripts of exactly what was said. If a participant is highly familiar with a given topic, or if a discussion outcome is exactly what they anticipated, then there is no need to record detailed information, if one or two carefully chosen words will suffice.

I don't usually write in sentences sometimes I just write one word that will be enough for me to remember what it was about.

People's roles also had an important effect on their note-taking. Managers tended to be involved in discussions around most agenda points which meant that they had fewer opportunities to take detailed notes. Note taking strategies were also influenced by whether or not the meeting was being minuted. Specifically, if participants knew that public notes were being taken they tended to take fewer notes.

these days I probably tend to take very few notes from meetings generally just things that are of importance to me or actions that I have taken out of a meeting, generally most important meetings would tend to be minuted anyway so I tend to rely on the minutes of the meeting.

We identified four main reasons why personal records were important to meeting participants:

- As personal reminders.
- To provide enough contextual information to carry out personal actions.
- To check the accuracy of the minutes.
- To brief others about what went on.

A major function of meetings is to agree on various actions that participants will carry out. Official minutes may contain insufficient contextual information to allow participants to carry out their actions. People therefore take notes to remind themselves about what actions they have committed to. The need for context about personal actions explains why personal notes tend to be esoteric and personalised; notes are intended to help participants carry out their own jobs rather than serving as a general public record. In other words, personal notes serve to record personal 'todo' items and their context, which participants fear they may otherwise forget:

if I failed to [carry out the action] immediately as time goes on it would start to slip out, there could be key points that I forget, or key actions that I forget to take. With a recorded note I can always check and make sure I've done them, or check what I have to do.

When no official minutes were taken of the meeting, then personal notes were sometimes shared among attendees to ensure that commitments were not forgotten:

I have so many meetings I would forget what happened if I didn't write them down. It is a memory aid for me and quite often it is a memory aid for other people at meetings so quite often other people will come to me and ask me what happened and I'll check my notes and see what I have written down.

Another important function of personal notes is to check the accuracy of the meeting minutes. All our participants reported using personal notes for this purpose. As the minutes are used both as the document of record and also as a group 'todo' list, participants were keen to ensure that they were

accurate, particularly about issues relating to themselves. For 25% of participants checking the minutes was the main function of their notes; after checking the minutes they discarded their own notes.

Finally, personal notes were sometimes used to report what went on in a meeting to non-attendees. However, when personal notes were taken to brief non-attendees, they tended to be less cryptic or personalised. Here note-takers felt they had to provide greater details of all aspects of the meeting that were thought to be relevant to the group being briefed.

The Limitations of Personal Records

Despite the value of personal notes, participants also complained about their limitations:

- Taking notes reduces one's ability to contribute to discussion.
- Personal notes sometimes lack both accuracy and comprehensibility.
- Their esoteric nature made them difficult for non-attendees to understand when they are shared.

One major problem was that taking accurate personal notes reduced participants' ability to participate in discussion. Participants' estimates of the time they took note-taking in meetings ranged from 5-40%, and all felt that this compromised their contributions:

if you are writing things down you are not listening to what is being said and it is probably more important to listen to what is being said rather than writing your own notes about things that have been discussed previously.

Indeed one of our informants pointed out that when he was chairing a meeting, he was so focused on the conversation and management of the meeting that he found it impossible to take notes.

This view is supported by our observations of informants. We noted down the frequency of note-taking and contributions to the meeting, and confirmed the expected negative impact of note-taking on people's contributions, with those taking detailed notes contributing least to the conversation

A second limitation of personal notes is that they were sometimes inaccurate or hard to interpret - even for those who had created them. One reason for this was the difficulty of simultaneously taking notes while listening to what is currently being said. Participants found themselves unable to process new information while writing detailed notes about an important prior point. The result was that personal notes could be cursory, disjointed and incomplete.

I can't understand my notes all the time probably because I have started to write down what I think I need to capture but then I have heard something else that has stopped me in my tracks, so what I have already written isn't joined up enough to understand what I was supposed to be capturing in the first place.

Others focused on trying to take fairly minimal notes, allowing them to contribute to and track the conversation, relying on their memories to reconstruct what went on. Again however there are limits to this strategy as such notes often weren't detailed or accurate enough to determine what went on in the meeting or what actions to undertake as follow up.

Technological Implications

These observations indicate clear problems with current techniques for the production and use of public and personal meeting records. We now use our user requirements data to critique current meeting browsers with regard to their support for public and personal recordings. For meeting browsers to be beneficial, they need both to support current record taking practices while addressing their main problems. Below, we assess current meeting browsers with regard to public and personal meeting records in turn.

Public Meeting Records

Current meeting browsers are highly focused on *single* meetings and as a result they do not generally support the collection of data from a *long-term series of meetings*. In cases where browsers allow access to multiple meetings (e.g. [1]) the user is required to search the entire meeting set in order to identify points of interest. There is little opportunity to perform a high level analysis on the meeting series; for example, tracking the progress of a task assigned in one meeting over a series of meetings.

Equally, the use of public meeting records as a record of past actions and decisions is not well supported by current meeting browsers. For example, public records represent meetings as a set of

decisions and actions, whereas current browsers support a restricted set of lower-level abstractions, only allowing users to search or browse for key words/phrases or by speaker.

There is also a question regarding the *formality* of the meeting record. Recall that one of the uses of public records is as a *contractual record* summarising the events of a meeting. It seems obvious, however, that the verbatim record provided by current browsers is too fine-grained to serve as an effective summary. In addition, the errorful nature of ASR makes it unlikely that participants would accept a formal summary of the meeting derived directly from ASR transcripts. It is also unclear whether hybrid transcript-centric access to the meeting record (such as supported in Whittaker et al. 1999,2002) would prove effective for this quasi-legal function. As suggested by Whittaker et al. (1994b, 1999), however, ASR-based browsers might be employed as a tool to help produce high quality meeting minutes. They could be used to clarify and identify key points and used to increase the efficiency of production and accuracy of meeting minutes with the desired formal status.

Although current meeting browsers have difficulty in matching some of the benefits of current public meeting records, they are able to overcome other problems associated with such records. Unlike minutes, automatically generated records are not laborious to produce, selective, or untimely. Furthermore, although there may be errors in the automatically generated annotations, the underlying recording is accurate, so that any inaccuracies in the annotations can be resolved by consulting the original recording. It is also possible to use browsers to determine contextual information relating to specific decisions, if the indices provided are suitable for locating main relevant points in the meeting. It is also easier to generate automatic records for all meetings. A dedicated minute taker is no longer required, allowing the minuter to increase their contribution to the meeting.

But current browsers do have problems relating to mobility and serendipity. Most meeting recording systems are designed for a specific room requiring extensive setup and calibration of recording equipment. Using a dedicated location precludes spontaneous meetings which account for over 90% of workplace interactions (Whittaker et al., 1994b). A novel approach to addressing this problem is outlined in (Lee et al., 2002), with the use of a portable recording device which allows for audio and video recording; whiteboard annotations, notes etc. are not included in this recording.

Finally, it is difficult to say whether current meeting browsers support the more peripheral experiences that our users said were missing from current public records. Audio-visual recordings should increase the experience of attending the meeting compared with a textual record. But novel browsers which aim to construct virtual meeting spaces (e.g. Cremers et al., 2005) may be required to provide an immersive meeting review environment.

Personal meeting records

Unlike public records, personal meeting records are less concerned with providing accurate, but abstract information for long-term analysis. The personalised nature of such records is not often reflected in today's meeting browsers, although mappings between personal notes and meeting records are addressed by some systems. Filochat (Whittaker et al., 1994b), Moran et al (1997) and Dynamite (Wilcox et al., 1997), for example, allow the user to take notes as they normally would, with these notes acting as an index into a recording of the meeting. Whilst notes-based systems are typically successful, current media rich meeting browsers do not yet exploit this.

Current meeting browsers might be used to provide contextual information for participants to check the accuracy of public records or to brief non-attendees about a meeting they had missed. However, current browsers are built around low level annotations (e.g. speaker turns, presented slides etc.) that do not readily support the extraction of personal actions, or the background information needed to execute these. Browsers which index personal notes taken during the meeting can support this process, but the support is inherently indirect. In Filochat, for example, there is no explicit way of qualifying the purpose of each note - the user must generate their own notation to achieve this.

With current meeting browsers, non-attendees can access the meeting record, so they do not have to rely on others' personal records to learn about a missed meeting. Again, however, there is a problem of abstraction - since a non-attendee has no means of quickly determining the salient points of a meeting; e.g. the actions that were assigned to them. One way to identify important events might be to analyse enhanced electronic records of group activity, e.g. collective note-taking or nonverbal indicators of participant interest.

But a significant problem with personal notes still remains; taking notes reduces the ability to participate in the meeting. Whilst current meeting browsers do not necessarily remove the need to take notes, they can significantly reduce that burden. And if personal notes are being used to construct an index into the meeting, rather than a complete summary of salient events, the note can be far more cursory allowing greater participation (Moran et al., 1997, Whittaker et al., 1994b).

Summary

Due to the reciprocal nature of public and personal meeting records, the current generation of meeting browsers both address some of the limitations of public records and replicate the benefits of personal records. The main failings of meeting browsers seem to be that they are largely only able to offer a view of a single meeting, that the main concepts of interest to users (namely decisions and actions) are not explicitly represented in the interface and finally that data collected from personal notes are not exploited in media rich meeting browsers. Furthermore, an unanswered question is whether a formal set of minutes are required now that users have access to the verbatim record.

Task-Based Browser Evaluation

The previous section identified user requirements for meeting records and browsers. We now take those requirements and use them to carry out a task-based evaluation of a state of the art browser.

Evaluated Technology

A typical meetings browser is shown below. It presents users with audio and video information recorded from the meeting which can be accessed directly using player controls. Various analyses are carried out on the speech including speaker identification. Speech is transcribed, and presented in a transcript which contains formatting information showing speaker identification (signalled using colour coding), along with the utterances from each speaker. The transcript used in this experiment is human generated and therefore contains no errors. Clicking on a particular speaker contribution in the transcript begins playing the audio and video related to that contribution. The system also shows a profile indicating overall contributions of each of the speakers, using the same colour coding. This representation can be scrolled and zoomed allowing users to form an impression of overall speaker contribution levels. Finally the system shows accompanying artefacts including presentations and whiteboard activities. Slides are temporally indexed so that selecting a specific slide accesses other data at that point in the meeting. Whiteboard events are presented as video streams and cannot therefore be used to directly index into the meeting. Our earlier technology review indicates this browser is typical of the current state of the art, offering random access to audio and video as well as access via semantic representations such as the speech transcript and meeting slides.

Figure 1 About Here

Target Tasks and Meeting Data

To evaluate the browser we asked 5 users to carry out 10 representative retrieval tasks on a recorded meeting. The meeting was part of the AMI corpus (Meeting 1008c), and is part of a series of semi-scripted project meetings in which a 4 person team try to develop a novel remote control, in particular focusing on the functional design requirements of the remote control. Present in the meeting are participants with the following roles, a manager, a marketing person, a user interface designer and an industrial designer. The meeting lasts 25 minutes.

We asked the following questions classified as to whether they require users to extract gist or specific facts from the meeting. Gist questions draw on information distributed throughout the meeting so that an overall understanding is required. With specific questions, information is localized, to a particular part of the meeting. The gist questions were based on the user requirements study which identified the need for accessing decisions, actions, project tracking as well as more peripheral aspects of the meeting

such as its general atmosphere. The factual questions were based on another study (Wellner et al., 2005) where we asked users to identify key factual events in several meetings.

Gist:

1. What topics/agenda items are being discussed during the meeting?
2. What is the status of the project?
3. Were any decisions made and if so, what were they?
4. Were any of the participants given action items or “todos” and if so, what were they?
5. What was the atmosphere in the meeting like?

Specific Facts:

6. Why was plastic eliminated as a possible material?
7. Describe the remote control design the “user interface designer” proposes.
8. Who attended the meeting and what are their roles?
9. What is the goal of the meeting?
10. What main contribution did Ed make to the meeting?

Users and Procedure

Users were English speaking and were staff or students at Sheffield University. We had hoped to recruit some of the original meeting participants to answer questions but this did not prove practical.

Users were first given a brief verbal tutorial to familiarise themselves with the browser, followed by few minutes to practice using it on some test questions about a short meeting. When they had satisfactorily answered these questions and were familiar with the browser features they began the experiment proper.

We did not provide participants with meeting minutes to help answer the questions. This is because (a) our research suggested that official minutes were often not available for meetings; (b) a key function of the technology is to remove the need to provide such minutes or to expedite the process of generating minutes.

They were given the 10 tasks above in random order. They were asked to answer them as accurately as they could using the browser. We imposed a time limit of 30 minutes for the whole experiment to stop subjects from simply playing the recording from beginning to end to answer each question. Subjects were also asked to use the ‘think aloud’ procedure - where they continuously describe their decisions and actions with the interface to the experimenter. The experimenter recorded noteworthy actions, asked users to explain these if they were not obvious, and kept a rough record of how long subjects took to carry out the tasks. Subjects were allowed to answer task questions out of sequence and also to alter their prior answers, if information they accessed later caused them to change their minds. This happened on several occasions when subjects discovered information late in the experiment that was relevant to an earlier question. Some subjects strategically deferred gist questions (requiring information that was distributed throughout the meeting) until the end of the session - when they had more exposure to the whole of the meeting.

Results

We scored subjects’ responses according to whether their answers were correct or not. They were scored against model answers and users were awarded partial credit for partial answers. Questions and

scores are shown in Table 3. Answers to the question about atmosphere are necessarily subjective, and are therefore scored as correct. They are excluded from the analysis.

	Task Type	User1	User2	User3	User4	User5	Mean	Standard Deviation
Topics/agenda items	Gist	0.5	0.5	0	0	1	0.4	0.37
Project Status	Gist	0.5	0	1	0	1	0.5	0.45
Decisions	Gist	0.25	0	0.25	0.25	0.5	0.25	0.16
Actions/Todos	Gist	0.5	1	1	0	0.75	0.65	0.37
Atmosphere	Gist	1	1	1	1	1	1	0.00
	Gist						0.45	0.34
Plastic eliminated	Specific	1	1	0.5	0.5	1	0.8	0.24
Remote control design	Specific	0	0.75	0.5	0.25	1	0.5	0.35
Attendees	Specific	0.75	0.75	0.75	0.25	1	0.7	0.24
Goal	Specific	1	1	1	0	1	0.8	0.40
Ed.'s contribution	Specific	1	1	1	1	1	1	0.00
	Specific						0.76	0.25
	Overall						0.62	0.29

Table 1: Performance on the Task-Centric Evaluation

The table shows that overall performance was not good. Although the meeting record was complete and contained the information needed to answer each question correctly, users only answered slightly more than half the questions correctly. There also seemed to be an overall difference between gist and specific questions, with users performing better on the specific questions. In addition, certain gist questions seemed to be very difficult, e.g. identifying decisions, and main topics. Other specific questions seemed to be easy, for example identifying Ed's contribution. We explore these differences in more detail below.

In contrast to their objective performance, all users expressed the opinion that they had performed well. They felt that they had a good overall grasp of what had happened in the meeting, and had answered the questions correctly. One potential reason for this discrepancy is subjects are *sampling* the meeting, which necessarily means that they never accessed certain parts of the meeting. Most subjects may therefore be unaware they had missed information relevant to a specific question, because this was material they had never accessed.

Observations and User Comments

All subjects used a strategy of focusing on the ASR transcript. Four began by playing audio and watching video, but quickly realised that this was an inefficient way to identify regions relevant to a given question. They therefore switched to using the transcript as a way to quickly identify relevant regions. This seemed to be because the transcript contains a high density of information whilst being easy to skim through, in a way the other data streams are not. Most users generally relied on the transcript alone for answering questions although two users towards the end of the experiment changed to consulting the audio to identify specific information they had already partially localised using the transcript. One subject stated that he thought that the audio and video were redundant. But despite its general utility, the transcript was problematic to scan, because its lack of formatting makes it hard to read (often ungrammatical text, no capitalisation, punctuation, large chunks of text with few breaks), and three subjects were negative about this. As a result, one subject switched to reading presentation slides as these were 'easier to read'. Another subject used a strategy of playing the meeting from beginning to end, while following the transcript simultaneously.

Four subjects attempted to sample what they thought would be significant regions of the meeting using their knowledge of meeting structure. When asked about the goals/main topics of the meeting they therefore focused on the beginning and end of the transcript, because this was where they expected the agenda to be discussed and decisions/actions reviewed.

Two subjects noted that the interface did not provide an abstract view of the meeting. This made it hard for them to find general information about the meeting or to navigate to particular points of interest. One subject specifically asked whether minutes were available (they were not), as these would have served as a useful overview and navigation tool. Two subjects used the slide presentation as an abstract

navigational aid, first finding a slide relevant to their current interest and navigating to the relevant part of the transcript and occasionally audio/video. Finally two subjects asked whether search could be provided, although it is not clear that this would have helped with all questions - in particular with some of the gist questions as people do not commonly mark decisions, actions or goals lexically.

All subjects felt that speakers should not only be colour coded but labelled by name too. A final complaint about the interface was it was poorly organised and 'too busy'. In support of this is the fact that one subject never commented on, or used the slides at all – possibly being distracted by the complexity of the display.

Design Implications

Media Presentation and Reduction of Interface Complexity

Video and audio do not seem to be the main focus for most subjects. On the whole subjects very quickly stopped playing the video and audio and opted to ignore these data streams. For most subjects the transcript provided the central means of browsing the meeting - sometimes supplemented with audio or slides. They rarely paid any attention to the video whilst carrying out the tasks, except when attempting to analyse the atmosphere of the meetings. Subjects found the audio useful but only in conjunction with the transcript, e.g. when they were able to accurately locate relevant sections to play using the transcript. In this way, the audio was used to address readability problems with the transcript. Although the transcript is a reasonable record of what was said in the meeting, it is difficult to quickly decipher large blocks of ungrammatical text.

Given their relatively minor importance, a clear design implication here is that audio and video should be less focal to the interface and be only accessible from the transcript (or some other interface abstraction). We are not suggesting they be removed from the UI as most of the subjects used *all of the data streams* to some effect during the experiment.

The Need for Abstraction

Subjects found it particularly difficult to answer gist questions, requiring an overall understanding of the meeting. For example, none of the subjects were able to locate the decisions or other high level information in the meeting without reviewing the meeting in its entirety. They were also unable to answer questions about the goal of the meeting or project status until they had listened to large chunks of the meeting, towards the end of the experiment.

This suggests the need for the interface to provide such abstract information. One user suggestion was to provide an interface resembling meeting minutes, which could provide both abstract information as well as other information provided very simply by minutes (e.g. names of participants) that was cumbersome to locate using the current browser. Other subjects used the transcript or the meeting slides as a way to gain a general overview of the meeting (although this was somewhat laborious). This suggests that there might be multiple ways to present such abstract information, including minutes, meeting summaries, topics, or marking decisions and actions as landmark indices. Subjects also attempted to infer their own abstractions of the data using implicit knowledge of meeting structure, for example four subjects sampled the beginning and end of the meeting on the grounds that this would give them an overview of what the entire meeting was about. However, this often meant that the user missed important information located in the middle of the meeting. This suggests it would be useful to provide users with landmarks showing where significant information such as decisions, actions, discussions of the agenda might be located.

The crucial need for an abstract representation of the meeting is also shown by the fact that users were unaware of the fact that they had missed important information. Despite performing poorly overall, all the subjects expressed the belief that they had gained a good insight into the contents of the meeting. If users had a clearer overview of the meeting contents, they should be more aware of the fact that they had yet to access significant information.

The Need for Context

When subjects incorrectly completed tasks, it was because they misunderstood the information (often derived from the transcript) due to lack of context. That is, they were not listening to entire discussions but dipping in and out of conversations and erroneously interpreting the information.

Many of the answers they provided were incomplete. Having found a piece of information to partially satisfy a task, they moved on to the next task. In fact, for open-ended questions, such as “what decisions were made in the meeting”, subjects were unable to verify that they had all the information pertaining to the task *unless they reviewed the meeting in its entirety*. The same was true for more specific questions too, where they required information about the local context.

One way to provide local context might be to support ways to rapidly drill down to access audio and video concerning a specific topic accessed from the transcript. Global context might be provided using meeting minutes or a summary showing where the user is currently focused in relation to the entire meeting and its key events. In other words, if the user is currently listening to decisions and wrap up at the end of the meeting, the fact that they are accessing the end of the meeting should be shown by highlighting the appropriate parts of the summary/minutes to show how the current information relates to the main events in the remainder of the meeting.

Conclusion

Overall the study shows that the data-oriented approach to browsing meetings embodied in current browsers does not effectively meet user requirements and that a more abstractive approach is needed. Specifically the browser is not fulfilling the function of traditional meeting records such as minutes; some of the most salient events from a meeting (such as decisions and actions), cannot be easily retrieved without reviewing the meeting in its entirety. The browser we tested presents the data in a way that facilitates reviewing from beginning to end, but users also need a view of the data that provides a more directed and less sequential review of the meeting. Many users tried to break away from sequential reviewing by concentrating on the beginning and end of the meeting, but this often meant they missed important information located in the middle of the meeting. There are two ways to counter-act sequentiality, the first is to provide the user with more and different index points in the data, and the second is to present the data according to some other high-level logical organisation.

Comparative Evaluation: Evaluating new technology for gisting meetings

Our final study addressed a major problem identified in the previous study; the lack of direct support for high level gist information. Here our focus was in trying to provide users with methods to rapidly identify the main points of a meeting.

We also employed a different evaluation that compared multiple browsing techniques. The previous study carried out requirements-centric evaluation of a single browser. We now describe a different evaluation technique that allows designers to compare multiple browsers. One problem with task-centric evaluation is that user tasks and the questions asked vary widely when people assess different browsers. Tasks and questions are often loosely defined, so final scores are open to considerable interpretation. In many other fields of research, an objective measure of system performance along with a standard corpus and set of reference tasks has been of enormous benefit in helping researchers compare techniques and make progress. For example, in the field of speech recognition, this has made possible the construction of real time, large vocabulary systems that would not have been feasible ten years ago. The text retrieval conference (TREC) has also used standard corpora, tasks and metrics with great success: average precision doubled from 20% to 40% in the last seven years.

One solution for evaluating meeting browsers is the BET (Browser Evaluation Test) (Wellner et al., 2005) which is a comparative technique for assessing how well browsers allow users to extract facts from meetings. One major limitation of the BET is that it makes the assumption that the key browser function is to *access facts* from the meeting. However, our requirements studies show that users also want to be able to extract gist from a meeting too. To this end, we have developed a modified version of the BET for evaluating how well a browser allows access to gist. We therefore report a study where we developed a technique for direct comparative evaluation of multiple browsers or browsing techniques. A second aim of the study was to gather detailed behavioural data about how different interface techniques are used to access meetings information. And given the lack of success of the complex browser we evaluated in the second study, we also wanted to evaluate the effectiveness of relatively simple interface techniques.

Examining Temporal Compression Methods

Our comparative experiment investigated the effectiveness of using *temporal compression* as a means of presenting the gist of meetings to users. Temporal compression is a technique which reduces the length of a speech recording but retains the important content. Reduction is primarily achieved by either removing portions of the original recording, or by speeding up the recording (Tucker and Whittaker, 2005). The temporal compression technique would seem best suited towards building a global understanding of the content of the recording rather than identifying specific facts contained within the recording. Temporal compression also relies on relatively simple interface techniques in contrast to the browser we evaluated in the previous study.

Our focus on gist means we cannot employ existing evaluation metrics that measure users' ability to extract specific factual information (Wellner et al., 2005). Instead we looked into evaluation metrics used in summarization. A common approach for assessing summarization algorithms is to have humans generate a summary (referred to as the 'gold-standard') and then measure the similarity of any automatic summary to this gold-standard.

While generating the gold-standard is relatively costly, once constructed it is a reusable resource that allows the automatic evaluation of different summarization algorithms (Mani, 2001, Nenkova and Passoneau, 2004). In our case however, whilst we can produce a similar human generated gold-standard summary to evaluate each temporal compression technique, we would still need listeners to manually produce a summarization of what they had heard. This would make any evaluation prohibitively time-consuming.

Because fact-finding methods are inappropriate and the summarization approach is too time-consuming, we devised a hybrid approach to capitalise on the advantages of both techniques. This process is illustrated in Figure 2. We produce a gold-standard ranking of utterances from the excerpt by selecting representative target utterances (see below) and asking human judges to rank these target utterances in order of their importance, in the excerpt.

To evaluate a temporal compression algorithm we compress the same excerpt and present it to human listeners. We then ask those listeners to rank the importance of the same subset of target utterances previously evaluated by the judges. If the temporal compression algorithm supports effective extraction of gist, we should find that the judges' and users' importance rankings are highly correlated. We can therefore compare different browsing techniques by seeing how closely they emulate the gold standard condition.

Assessment Procedure

Given the general procedure described above, the goal of the experiment was to objectively compare different excision and speedup approaches to producing temporally compressed audio. We also wanted to know whether these techniques were better than uncompressed playback in allowing users to identify the gist of the recording, and how people used compression techniques when they were allowed active control over their browsing.

Our algorithms derived insignificant utterances and words from human generated transcripts. But in many practical situations these would not be available, and our algorithms would have to rely on errorful ASR-generated transcripts, where exact error rates depend on many factors including recording conditions, ASR algorithm, language and acoustic models. In this study we did not use such errorful transcripts, because we wanted to test the upper bound of excision techniques.

This section provides more details about the procedure we used.

Preliminary Preparations

Figure 2 About Here

We chose 36 four-minute and 8 thirty-minute meeting excerpts from the ICSI meeting corpus, using manual transcripts supplied with the corpus. Using these transcripts we measured the importance of

words using simple information retrieval measures of term frequency * inverse document frequency (TFIDF) (Sparck-Jones, 1972).

We then derived the importance of each utterance in the transcript by computing the mean importance of all words contained in each utterance. We chose target utterances that were highly important, unimportant and a number with intermediate levels of importance, ensuring that each was meaningful and non-repetitive. Target utterances were less than a minute long, representing speech from a single speaker, they were a mean of 16 words in length.

Constructing the Gold-Standard

We then built a small web-based application to collect the judges' gold-standard target utterance rankings. Other research reports that judges experience problems in ranking collections of spoken utterances, but make more accurate judgments when presented with transcripts of the speech (Walker et al., 2004). We therefore allowed judges to read transcripts of the recordings and rank the importance of target utterances on-line. They were given one month to complete their rankings. Fifteen judges were used, meaning that we collected three independent rankings for each meeting excerpt. Judges had an unlimited amount of time to rank utterances.

To ensure that judges were in agreement we measured Kendall's coefficient of concordance for the rankings provided for each meeting. The concordance coefficient in each case was greater than 0.6, with mean concordance 0.75, indicating agreement ($p < 0.05$). We then constructed the gold-standard rankings by computing the mean ranking for each target utterance across judges, assuming that rankings are linear. Note that this means that target utterances can be assigned non-integral ratings

Compression Techniques

We used three different algorithms to produce the compressed excerpts: utterance level excision, word level excision and a heuristically motivated speedup algorithm. These were compared with a non-compressed baseline.

Insignificant Utterance Excision

There are many possible ways to implement utterance excision (Mani, 2001, Nenkova and Passoneau, 2004). We used a simple method that did not require complex natural language or acoustic processing. We first computed utterance importance scores using the TFIDF measures described above. We then

used these scores to rank utterances according to their importance. To ensure that the compressed recording contained all the target utterances selected for this excerpt, we began with an audio file that included each of our target utterances. This meant that users are not biased by being asked to rank utterances that they did not hear. We then progressively included high-ranking utterances until the file was of the duration determined by the compression amount. Utterances were presented in the order in which they occurred in the original recording.

Insignificant Word Excision

This method is identical to utterance excision, except that importance scores are calculated for each word and high ranking words added to the file containing target utterances until the desired level of compression is achieved. We removed stop words from all target utterances prior to producing the compressed excerpt, so that target utterances would not appear unusual, when users heard them. The following example shows the effects of excision:

Speed-Up

We used the mach1 speedup algorithm (Covell et al., 1998) which aims to replicate the phonetic speed variations which occur when humans naturally modify their speech rate. We first compute a measure of the relative speech rate for each part of the recording. We then linearly transform this relative speech rate contour to reach the desired level of compression. This transformed contour is then used to dynamically alter the speech rate using a standard SOLA algorithm (Verhelst, 2000).

Figure 3 About Here

Control: Non-Compressed Excerpts

Uncompressed control excerpts allow us to directly compare performance between compressed and uncompressed excerpts.

Compression Levels

In addition to modifying the *type* of compression we also modified the *level* of compression to see whether a technique's effectiveness depends on the amount of compression. This allows us to tell, for example, whether speedup works at low compression levels but is ineffective at higher ones. The levels of compression were chosen to reflect levels beginning at those cited as being comfortable for listeners (Arons, 1997). We also used long and short extracts expecting the benefits for temporal compression to be greater for longer stretches of speech. We also manipulated the amount of control that users had over what they heard. In the long condition we gave them a simple interface (see Figure 3) allowing them to control what they heard, replay parts that they had misunderstood or uncompress parts of the speech that were hard to grasp. In the short condition they had no control over what they heard; they simply listened to clips from beginning to end. For the short excerpts we applied three levels of compression (66, 50 and 40% of the original duration, corresponding to 1.5, 2 and 2.5 times normal speed) and for the longer excerpts two levels of compression were applied (66 and 50% of the original length).

Users

Twelve users were selected from university staff and students. They were aged between 20 and 40. None reported any hearing difficulties and each was paid for their participation.

Experimental Procedure

Users attended experiments over four days, and were presented with a single compression technique each day. Each experiment contained two phases, the first phase involved short continuous excerpts and the second phase longer excerpts where people used the browser to control what they heard.

In the first phase, users heard nine different compressed excerpts (three repetitions of each compression rate). After hearing each complete excerpt, they were presented with the set of target utterances from the excerpt, and asked to rank the importance of each target utterance. Users performed their ranking by choosing from five labels ("important" to "unimportant") from drop-down menus next to the target utterances. The target utterances were presented in a random ordering to each user.

Before carrying out the second phase, we gave users a short web-based tutorial that explained the functioning of the browsing interface. (shown in Figure 3). The tutorial explained each interface feature (e.g. pause, backup, uncompress, replay), encouraging listeners to practice using that feature. They were allowed as much time as they liked to familiarize themselves with the interface before proceeding to the next part of the task. Interface practice was provided each day as different compression techniques may demand different usage strategies.

In the second phase users had thirty minutes to listen to each excerpt using the browser. We imposed a time limit as we were interested in efficient processing of speech. We chose 30 minutes as the enforced time limit as this was the uncompressed duration of the excerpt. The experiment finished either after thirty minutes elapsed, or the user decided they had fully processed the excerpt, whichever happened sooner (typically users finished before the thirty minute deadline, except in the uncompressed condition). To keep users aware of their progress, the interface included a timer showing how much time remained for them to complete the task. It was made clear that users needed to leave themselves enough time to listen to the entire recording. They were then presented with the twenty target utterances and were asked to independently judge the importance of each of the utterances. Judgements were made with the same ranking labels as before but no restrictions were placed on how many times each label could be applied to an utterance. Users had unlimited time to perform their rankings or judgements in both phases.

Performance Measures and Data Collected

We used Kendall's tau to measure the agreement between the judges' gold standard rankings and the user judgements. Close agreement indicates a successful browsing technique. This simple agreement measure does not, however, capture a key aim of temporal compression - which is to reduce the time taken to effectively process a recording. Thus we normalise the tau measure (by the mean tau level for the relevant condition, across all subjects and compression types), and then divide this by the normalised time taken (the listening time divided by the original length of the clip – note that the listening time was fixed for the short condition). We denote this score the *comprehension efficiency*. In this case if the actual tau is equal to the average tau, then no compression would yield a score of 1 for efficiency, and using half the original time an efficiency score of 2, etc. In this way it should be possible to observe how time savings trade-off against gist scores.

For the short clips the median tau value was 0.60 (there is significant skew because of the small number of sentences that were ranked), and for the long clips the mean tau value was 0.31. The relatively low value for the long clips does not impact our subsequent analysis since we are comparing the level of understanding across compression types, but it is indicative of the difficulty of extracting gist from just the audio of such lengthy recordings (note that the judges had both infinite time and the ability to assess the target utterances within the context of the transcript).

We also collected subjective data at the end of each phase of the experiment, examining both users' perception of the techniques and, in the second phase, their perceived use of the interface. The phrases used for Likert responses in both phases were: "I found it easy to get the gist of the discussion", "I felt I was missing important information" and "I felt the speech was too fast". However in the second phase we also used the phrases: "I repeatedly had to go back in the speech", and "If I missed something I could easily find out what it was".

In the second phase we also logged all user interface actions, their timing and duration. This enabled us to determine how frequently users paused, rewound and replayed each excerpt, as well as whether they were listening with or without compression.

Hypotheses and Results

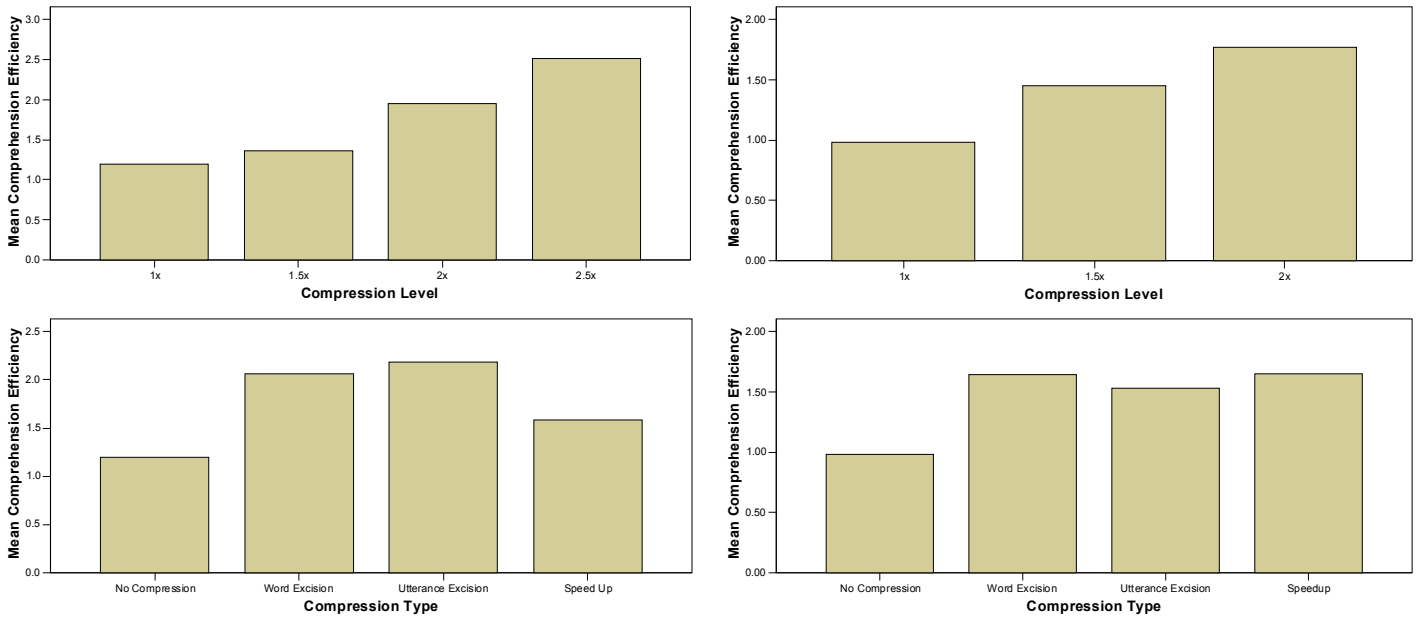


Figure 4. Bar graphs showing comprehension efficiency against compression level (top) and compression type (bottom) for the short (left panel) and long (right panel) excerpts.

For both short (S) and long (L) excerpts, we made predictions about compression level, type and subjective reactions to different compression types, and about how people would actively use the browser to control different types of compression.

Efficiency Hypotheses

S1/L1: Compression Level

Greater compression should allow users to identify gist more efficiently. We therefore expected greater comprehension efficiency at higher compression levels for all compression techniques.

To assess the efficiency hypotheses we conducted a 4 (compression level) X 4 (compression type) ANOVA for the short excerpts and a 3 (compression level) X 4 (compression type) ANOVA for the long excerpts. In both cases the dependent variable was comprehension efficiency. As expected comprehension efficiency is increased at higher compression levels.

Figure 4 (top left and right) shows the relationship between comprehension efficiency and compression level for both long and short excerpts. ANOVAs confirmed the effect of compression level on comprehension efficiency ($F(2,422)=12.43, p<0.01$; $F(1,89)=16.00, p<0.05$, for the short and long excerpts respectively). There is a significant correlation between comprehension efficiency and compression level ($r(432)=0.21, p<0.001$ (short); $r(96)=0.47, p<0.001$ (long)). This shows that users are able to extract gist even at the highest levels of compression. In other words for the long excerpts, participants are able to save approximately 13 minutes listening time while obtaining equivalent gisting levels to the uncompressed case.

S2/L2: Compression Type

A previous study (Tucker and Whittaker, 2005) led us to expect differences between compression techniques. Specifically we predicted that excision techniques would be more efficient than both speedup and uncompressed speech.

The results confirm the superiority of excision techniques over speedup for short excerpts. There were no differences for long excerpts. The mean comprehension efficiency organized by compression type is shown in Figure 4 (bottom left and right). The ANOVAs indicate that an effect of compression type is only present for short excerpts ($F(2,422)=3.84$, $p<0.05$, for short, $F(1,89)=0.52$, $p>0.6$, for long). Tukey planned comparisons here indicate, as we predicted, that the excision techniques outperform speedup for short excerpts (both $p<0.05$). There were no interactions between level and type of compression indicating that the relative success of a technique did not depend on the amount of compression.

Subjective Hypotheses

S3/L3: Subjective Responses

We expected users' subjective responses to replicate the findings of our initial study (Tucker and Whittaker, 2005), showing a preference for excision over speed up.

Excision was perceived as generally better than speedup for short excerpts, with speedup perceived as 'too fast' for long excerpts. We conducted a second set of ANOVAs, with the appropriate subjective ratings as dependent variables (see Figure 5).

For short excerpts, the ANOVA indicated an effect of compression type on each of the 3 subjective responses ($F(3,44) = 12.64$, $p<0.05$; $F(3,44) = 8.42$, $p < 0.05$; $F(3,44) = 31.265$, $p < 0.05$). Users found it hardest to 'extract gist' with speedup, although this difference was only a trend when comparing word excision and speedup. A similar set of responses was seen for the 'missing information' question. For 'I felt the speech was too fast', users stated that the sped up speech was too fast, compared with both excision conditions.

L4: Interface Responses

Our previous work indicated that speedup can present information too fast. We therefore expected users to report increased use of the rewind functions in the speedup case.

For the long excerpts, all techniques were judged the same for all questions except for the 'I felt the speech was too fast'. Here again sped up speech was judged as too fast compared with excision conditions and uncompressed speech ($F(3,44) = 10.87$, $p<0.05$; Tukey post hocs are all $p < 0.05$).

Figure 5 About Here

We also examined users' impressions of how they used the browser to address processing problems. While users felt they could recover missing information equally well in each compression condition, the ANOVA confirmed an effect of compression type on the response for 'I repeatedly had to go back in the speech'. Planned comparisons indicated that backing up was felt to occur most in the speedup condition (Tukey, $p<0.05$).

Interface Interaction Hypotheses

We also predicted how people would make use of the browser in the different compression conditions.

P1: Reactions in the Speedup Condition

A compression technique such as speed up induces cognitive overload. We therefore expected that speedup would lead users to pause or switch off compression to reduce this.

P2: Reactions in the Utterance Excision Condition

In contrast, utterance excision omits material, which introduces discontinuities in what users hear. They may understand the utterance they have just heard but have no context for it. We therefore expected users to be more likely with utterance excision to backup and replay an excerpt, to establish the context they require.

P3: Reactions in the Word Excision Condition

We expected word excision to occupy the middle ground between the two other compression techniques. Removing a large number of words may both reduce comprehensibility of the current utterance as well as leading to a loss of context for subsequent utterances. Because of this, we predicted that users would be more likely to both turn the compression off and rewind the recording in this condition.

To analyse how the browser was used in the different compression conditions, we clustered logfile entries into three categories according to problems users had described in our initial study. These categories were uncompress (in response to information generally being presented too fast), replay (in response to missing some information), as well as combined replay-uncompress actions (in response to being unable to understand what one has just heard). The mean number of these actions organized by the compression algorithm is shown in Figure 5.

We found different usage patterns for the different compression techniques: with uncompress being most used with speedup and back with utterance excision. As we predicted (P1), use of the uncompress button depends on compression type ($F(2,89)=4.76, p<0.01$), with planned comparisons showing this is more frequent in the speedup case than other conditions (all Tukey, $p<0.05$). Consistent with prediction P2, use of the back button differs for different compression types ($F(2,89)=3.46, p<0.05$). Planned comparisons showed that the back button is most used in the utterance excision condition compared with other conditions (all Tukey, $p<0.05$). Contrary to prediction P3, combinations of back-uncompress actions occur independently of the compression techniques ($F(2,89)=0.04, p > 0.9$).

Informal Comments

Users' posthoc informal comments mirrored the experimental data. People were usually positive about utterance excision, ('This was all very clear'), although they noted that the discontinuities it introduced could mean that they lacked context for what they were just hearing. ('[I] found the flow was disrupted when I was listening to a sentence...and it then skips onto another one'). This comment also indicates that users were sensitive to exactly how the different techniques were presenting information. They were slightly less positive about word excision, feeling it presented information as a 'stream of words' and that this could make it hard to understand, even increasing the perceived speed at which information was presented. 'The stream of words came so quickly at times that understanding became difficult'. People were much less positive about speedup, however, saying: 'The speech was too fast' and 'Important information...was simply too garbled'. There is some evidence that there is a learning effect for processing sped up speech (Arons, 1997) but we were unable to assess this in our experiments.

Summary and Discussion

We developed a method for comparing various new interface techniques intended to help users extract the main points of a recording, without having to rely on feature rich visual displays. We applied this method to compare simple but novel excision techniques that allow users to extract the gist of meetings. We found that excision techniques were generally better than speedup. They also outperformed an uncompressed baseline. Users also preferred excision techniques to speedup, and were less likely with excision to turn off compression. We did not find predicted differences between excision and speedup for the long excerpts, but potential differences may be masked by active user browsing, e.g. turning off speedup for the long excerpts. Finally lexical excision was surprisingly successful given the simplicity of the technique.

We also analysed users' behaviours to infer how the different techniques affect comprehension. Although excision techniques were generally successful, excision and speed up affect comprehension in different ways. Speedup can induce general overload, leading users to uncompress incoming material without replaying what they have heard, whereas excision can lead to context-loss, with the need to replay recently heard materials.

Overall these results are striking because they show that relatively simple interface techniques can support an important user task, gist extraction. Although extraction of gist may be more straightforward than some other information extraction tasks, our results are in contrast to our previous evaluation where a more complex interface performed poorly on a variety of different user tasks.

Overall Summary and Conclusions

This paper explored two related hypotheses for the failure of ICR systems. There is evidence to support our first hypothesis, that ICR systems have generally ignored user needs, with the result that existing systems do not address key user problems. Our survey showed that ICR research has predominantly focused on technology rather than on user-centric issues, and there is a dearth of studies of user needs. As a result, while many interesting systems have been built, these currently do not seem appropriate for users' ICR needs, and this may explain the lack of uptake for the systems that have been built.

Our own empirical studies reinforce this view, with a task-based evaluation showing that a complex state of the art browser did not satisfy key user needs we had identified. It failed to provide an appropriate level of abstraction to allow users to strategically focus on important parts of the meeting. Furthermore it did not support access to information categories that are important to users, e.g. summaries, agenda items, decisions and actions. That study also suggested that serious UI work is needed to better organise and focus interface elements on concepts that users find important. Our second study, in contrast, showed that relatively simple interfaces can support an important user need, namely the effective extraction of gist. Users were able to identify gist using interface techniques that selected the important parts of meetings, excising the unimportant ones. The second study suggests that simple designs that are focused on genuine user needs may be more effective than complex technologies developed without a focus on user requirements.

In order to improve current interfaces and to move beyond providing only rudimentary support for ICR, we also need to develop new methods for evaluating interfaces in order to determine how they should be modified to better support users needs. We explored two different methods to evaluate browsers, each of which has strengths and weaknesses. Task-oriented evaluation generates rich information about the quality of a specific browser. In our task-oriented evaluation, we were able to identify key weaknesses of a state of the art browser, and generate recommendations about how that browser might be improved to address its shortcomings. But the utility of task-oriented evaluation is dependent on ensuring that representative evaluation tasks are used and tested on a browser that is also representative of the state of the art (Whittaker et al., 2000). In contrast, the comparative method we used to evaluate temporal compression allowed us to compare multiple browsers and find out which was most effective for extracting gist. We were also able to collect fine-grained information about how different temporal compression algorithms were used in practice. But the comparative method also has weaknesses. Although it imposes greater control over the tasks that users executed, it offers more impoverished information about the ways in which evaluated interfaces can be improved to better support user access.

Our review of current ICR systems also suggests a number of outstanding technical issues. In particular achieving high quality audio records remains a critical problem. Many studies, including our task-oriented evaluation, show that ASR transcripts play a critical role in browsing and search. Without high quality recordings, speech recognition will be poor - with resulting ASR transcripts containing multiple errors thus compromising user access (Whittaker et al., 1999, 2002). Furthermore, recording technology needs to become more lightweight; ICR technology will not become pervasive until techniques are developed to record meetings in multiple settings. Current systems generally rely on dedicated meeting rooms - which necessarily reduces the settings in which meetings are recorded. This is a critical limitation; studies of workplace communication show that prearranged meetings are the exception rather than the rule, with the majority of communications being impromptu (Whittaker et al., 1994a). Until we can provide lightweight ways to capture meetings, in many different settings, ICR technology will necessarily not become pervasive. We need to develop systems similar to that of Lee et al. (2002) which aim to support ICR in multiple contexts.

There are also a number of other technical developments that may prove to be significant over the next few years. These include attempts to incorporate technologies into existing workplace artefacts by applying ideas from ubiquitous computing, to make technology less intrusive and easier to integrate into current meeting practices (Streitz, et al., 1998, CHIL, Yu et al., 2000). Other new developments include machine learning and other AI-based techniques being used to identify high-level categories or landmarks in the recording, such as 'hotspots', summaries, or conversational topics. However to realise their full potential, these new concepts need to be tied closely to careful interface development. Current interfaces are too complex for users, and adding yet more new features is unlikely to be a direct improvement unless these are carefully integrated into new systems in ways that support user needs.

There are also sociotechnical issues that have to be addressed before systems will become acceptable. For example, in our early user studies (Whittaker et al., 1994b) users expressed concerns about privacy, in particular about the effects that being recorded will have on the conduct of work conversations, and on the uses to which such recordings might subsequently be put. It was clear from these studies that users need to be provided with more control over recordings, including ways to quickly browse and remove potentially compromising speaker contributions or parts of the conversation that should remain off-record (Whittaker et al., 1994b). Developments in speech browsing and editing may be critical to addressing privacy concerns, by allowing users to easily identify and remove contentious portions of the meeting record (Whittaker et al., 1999, Whittaker et al., 2002, Whittaker and Amento, 2004). Our early studies also found that it is important to elicit views about privacy in the context of a concrete working system. Users' initially negative reactions to being recorded were substantially modified when they had extended experience with using a working system - in particular when they had clearer ideas about the benefits that can accrue to having meeting records readily available (Whittaker et al., 1994b, Whittaker and Amento, 2003).

A related point is that working ICR systems may lead to unanticipated changes in business practice. For example, it is not current business practice to record meetings, nor is it standard in meetings to access information from prior meetings, in order to build on conversations that took place in those prior meetings. Nor is it standard to construct minutes from recordings, or to use such recordings in lieu of minutes, partly as we have argued because recordings lack the formality that meeting minutes currently possess. But new technical possibilities may mean that organisations rethink their policies on conversational records and how these are used. New technologies may also lead to changes in meeting attendance. For example if users can easily access specific parts of a recorded meeting, the practice of 'auditing meetings' may become more common, i.e. listening to key parts of a meeting record after the event, rather than attending the meeting itself.

Finally ICR technology should have significant impact on theories of human communication. The widespread availability of recordings should make it possible to develop better theories about human communication and collaboration. ICR technology will make rich new datasets available for theorists of communication, but also facilitate research concerning important related questions such as memory for discourse or memory for collaborative activity.

References

- AMI project <http://www.amiproject.org>.
- Arons, B. (1997) SpeechSkimmer: A System for Interactively Skimming Recorded Speech. ACM Transactions on Computer-Human Interaction, 3-38
- Bett, M., Gross, R., Yu, H., Zhu, X., Pan, Y., Yang, J., Waibel, A. (2000) Multimodal Meeting Tracker. In: Proceedings of RIAO, Paris, France.
- Brotherton, J. A., Bhalodia, J. R., Abowd, G. D. (1998) Automated Capture, Integration and Visualization of Multiple Media Streams. In: Proceedings of The IEEE International Conference on Multimedia Computing And Systems, 54-63.
- CHIL Project, <http://chil.server.de/servlet/is/101/>
- Chiu, P., Boreczky, J., Girgensohn, A., Kimber, D. (2001) LiteMinutes: An Internet-Based System For Multimedia Meeting Minutes. In: Proceedings of 10th WWW Conference, Hong Kong, 140-149.
- Christel, M.G., Smith, M.A., Taylor, C. R., Winkler, D.B. (1998) Evolving Video Skims Into Useful Multimedia Abstractions. In: Proceedings of CHI '98, Los Angeles, CA, 171-178
- Colbath, S., Kubala, F., Liu, D., Srivastava, A. (2000) Spoken Documents: Creating Searchable Archives From Continuous Audio. In: Proceedings of 33rd Hawaii International Conference On System Sciences.

- Michele Covell, Margaret Withgott, and Malcolm Slaney (1998) Mach1: Nonuniform Time-Scale Modification of Speech, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Seattle, WA. May 12-15.
- Cremers, A.H.M, Hilhorst, B. and Vermeeren, A.P.O.S. (2005). "What was discussed by whom, how, when and where?" personalized browsing of annotated multimedia meeting recordings, HCI International, Las Vegas USA.
- Cutler, R., Rui, Y., Gupta, A., Cadiz, J.J., Tashev, I., He, L., Colburn, A., Zhang, Z., Liu, Z., Silverberg, S. (2002) Distributed Meetings: A Meeting Capture And Broadcasting System. In: Proceedings of 10th ACM International Conference on Multimedia, Juan-les-Pins, France, 503-512
- Degen, L., Mander, R., Salomon, G. (1992) Working With Audio: Integrating Personal Tape Recorders And Desktop Computers. In: Proceedings of CHI '92, Monterey, CA, USA, 413-418.
- Ferret Browser, <http://mmm.idiap.ch/>
- Flynn, M., Wellner, P.D. (2004) In Search of a Good BET: A Proposal for a Browser Evaluation Test. IDIAP IDIAP-COM03-11.
- Foote, J., Boreczky, G., Wilcox, L. (1998) An Intelligent Media Browser Using Automatic Multimodal Analysis. In: Proceedings of ACM Multimedia, Bristol, UK, 375-380.
- Geyer, W., Richter, H., Fuchs, L., Frauenhofer, T., Daijavad, S., Poltrock, S. (2001) A Team Collaboration Space Supporting Capture And Access Of Virtual Meetings. In: Proceedings of 2001 International ACM SIGGROUP Conference On Supporting Group Work, Boulder, Colorado 188-196.
- Girgensohn, A., Borczyk, J., Wilcox, L. (2001) Keyframe-based User Interfaces For Digital Video. IEEE Computer, 61-67.
- Hindus, D., Schmandt, C. (1992) Ubiquitous Audio: Capturing Spontaneous Collaboration. In: Proceedings of 1992 ACM Conference on Computer-Supported Cooperative Work, Toronto, Ontario, Canada 210-217
- IM2 Project, <http://www.im2.ch/>
- Janin, A. et. al (2004) The ICSI meeting project: Resources and research, in Proc. of ICASSP 2004 Meeting Recognition Workshop.
- Kazman, R., Al-Halimi, R., Hunt, W., Mantei, M. (1996) Four Paradigms for Indexing Video Conferences. IEEE Multimedia 3(1), 63-73.
- Kazman, R., Kominek, J. (1997) Supporting the Retrieval Process In Multimedia Information Systems. In: Proceedings of Proceedings of the 30th Annual Hawaii International Conference On System Sciences, Hawaii, 229-238.
- Kimber, D.G., Wilcox, L.D., Chen, F.R., Moran, T.P. (1995) Speaker Segmentation For Browsing Recorded Audio. In: Proceedings of CHI '95, 212-213.
- Lalanne, D., Sire, S., Ingold, R., Behera, A., Mekhaldi, D., Rotz, D. (2003) A Research Agenda For Assessing The Utility Of Document Annotations In Multimedia Databases Of Meeting Recordings. In: Proceedings of 3rd International Workshop on Multimedia Data And Document Engineering, Berlin, Germany.
- Lee, D., Erol, B., Graham, J., Hull, Jonathan J., Murata, N. (2002) Portable Meeting Recorder. In: Proceedings of ACM Multimedia, 493-502.

M. Mantei (1988) Capturing the Capture Lab Concepts: A Case Study in the Design of Computer Supported Meeting Environments. In Proceedings of the Conference on Computer Supported Cooperative Work, Portland, Oregon, September 1988.

M4 Project, <http://www.m4project.org/>

Mani, I. (2001) Summarization Evaluation: An Overview. In *proceedings of the NTCIR Workshop*, 2001.

Moore, D. (2002) The IDIAP Smart Meeting Room. IDIAP-COM 02-07, November 2002.

Moran, T., VanMelle, W., Chiu, P. (1998). Spatial interpretation of domain objects integrated into a freeform electronic whiteboard, UIST, 175-184.

Moran, Thomas P., Palen, L., Harrison, S., Chiu, P., Kimber, D., Minneman, S., Melle, W., Zellweger, P. (1997) "I'll get that off the audio": A Case study of salvaging multimedia meeting records. In: Proceedings of CHI '97, Atlanta, Georgia, 202-209.

N. A. Streitz, J. Geisler, and T. Holmer (1998) "Roomware for Cooperative Buildings: Integrated Design of Architectural Spaces and Information Spaces," Cooperative Buildings: Integrating Information, Organization, and Architecture, Springer-Verlag, Lecture Notes in Computer Science, 1370, pp. 4-21.

Nenkova, A and Passoneau, R. (2004) Evaluating content selection in summarization. In Proceedings of the HLT-NAACL conference.

Olson, J.S., Olson, G.M., Storrøsten, M., and Carter, M. (1992) How a Group-Editor Changes the Character of a Design Meeting as well as its Outcome, in Proceedings of CSCW'92, pp 91-98.

Pedersen, E., McCall, K., Moran, T.P., and Halasz, F. (1993). Tivoli: An Electronic Whiteboard for Informal Workgroup Meetings. Proc. of InterCHI 1993, 391-398.

Poole, M.S. & DeSanctis, G. (1989). Use of group decision support systems as an appropriation process. HICSS Conference, 1989, 149-157.

Roy, D.K., Schmandt, C. (1996) NewsComm: A Hand-Held Interface for Interactive Access To Structured Audio. In: Proceedings of CHI '96, 173-180.

K. Spark-Jones (1996) A Statistical Interpretation of Term Specificity and its Application in Retrieval, Journal of Documentation, 28, pp 11-21.

Tucker, S., and Whittaker, S. (2005). Novel techniques for time compressing speech: an exploratory study. In IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, USA.

Tucker, S., Whittaker, S (2004) Accessing Multimodal Meeting Data: Systems, Problems and Possibilities, In Proc. of MLMI'04, Springer-Verlag, 1-11.

Verhelst, W. (2000) Overlap-Add Methods for Time-Scaling of Speech, Speech Communication, volume 30, No. 4, pages 207—221.

Wellner, P., Flynn, M., Guillemot, M (2004) Browsing Recorded Meetings With Ferret, In Proc. of MLMI'04, Springer-Verlag, 12-21.

Wellner, P., Flynn, M., Tucker, S., and Whittaker, S. (2005). A Meeting Browser Evaluation Test. In Proceedings of CHI05 Conference on Human Factors in Computing Systems, New York: ACM Press.

- Whittaker, S. and Amento, B. (2003). Seeing what your are hearing: Co-ordinating responses to trouble reports in network troubleshooting. In *Proceedings of European Conference on Computer Supported Cooperative Work*, 219-238. Kluwer, Netherlands.
- Whittaker, S., and Amento, B. (2004). [Semantic Speech Editing](#). Proceedings of Conference on Computer Human Interaction, 527-534, New York, ACM Press.
- Whittaker, S., Davies, R., Hirschberg, J., and Muller, U. (2000). [Jotmail: a voicemail interface that enables you to see what was said](#). In Proceedings of CHI2000 Conference on Human Computer Interaction, 89-96. New York: ACM Press.
- Whittaker, S., Frohlich, D., and Daly-Jones, O. (1994a). [Informal communication: what is it like and how might we support it?](#) In Proceedings of CHI'94 Conference on Computer Human Interaction, Boston, USA, Eds., C. Plaisant, NY: ACM Press, 130-137.
- Whittaker, S., Hirschberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., Stead, L., Zamchick G., & Rosenberg, A. (2002) [SCANMail: a voicemail interface that makes speech browsable, readable and searchable](#). In Proceedings of CHI2002 Conference on Human Computer Interaction, NY: ACM Press, 275-282.
- Whittaker, S., Hirschberg, J., Choi, J., Hindle, D., Pereira, F., and Singhal, A. (1999). [SCAN: designing and evaluating user interfaces to support retrieval from speech archives](#). In Proceedings of SIGIR99 Conference on Research and Development in Information Retrieval, 26-33.
- Whittaker, S., Hyland, P., Wiley, M. (1994b) Filochat: Handwritten Notes Provide Access To Recorded Conversations. In: Proceedings of CHI '94, Boston, Massachusetts, USA 271-277
- Whittaker, S., Terveen, L., and Nardi, B. (2000). [Let's stop pushing the envelope and start addressing it: a reference task agenda for HCI](#). Human Computer Interaction, 15, 75-106.
- Wilcox, L., Schilit, W., Sawhney, N. (1997) Dynamite: A Dynamically Organized Ink and Audio Notebook . Proceedings of CHI '97, March 1997, 186-193.
- Yu, H., Tomokiyo, T., Wang, H., Waibel, A. (2000). New Developments In Automatic Meeting Transcription (2000) In Proceedings of ICSLP, Beijing, China.
- Walker, M., Whittaker, S., Stent, A., Maloor, P., Moore, J., Johnston, M., Vasireddy, V. (2004). [Generation and Evaluation of User Tailored Responses in Dialogue](#). Cognitive Science, 28, 811-840.

Figure Captions

Figure 1 – A screenshot of a state of the art meeting browser

Figure 2 – A schematic of the comparative evaluation process

Figure 3 – A screenshot of the simple browser used in the experiments

Figure 4 - Bar graphs showing comprehension efficiency against compression level (top) and compression type (bottom) for the short (left panel) and long (right panel) excerpts.

Figure 5 – Interface Strategies for the different Compression Conditions