IDIAP RESEARCH REPORT

WHEN USERS MEET TECHNOLOGY: THE MEETING BROWSER DEVELOPMENT HELIX

Andrei Popescu-Belis Denis Lalanne
Hervé Bourlard

Idiap-RR-05-2011

MARCH 2011

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

When Users Meet Technology: The Meeting Browser Development Helix

Andrei Popescu-Belis
Idiap Research Institute
Martigny, Switzerland
andrei.popescu-belis@idiap.ch

Denis Lalanne
Department of Computer Science
University of Fribourg, Switzerland
denis.lalanne@unifr.ch

Hervé Bourlard
Idiap Research Institute
Martigny, Switzerland
herve.bourlard@idiap.ch

March 7, 2011

Abstract

This paper shows how the task of assistance to fact-finding has gradually become central to the field of meeting browsing. Requirements elicitation studies could not lead alone to a precise specification, because they depend on the preliminary assumptions of each study. Therefore, user studies were gradually focused towards the most promising task, namely fact finding or verification in multimedia meeting recordings. This task answers significant user needs, has enough theoretical interest, and is within reach of current technology, as illustrated by a variety of meeting browsers, including end-user products for conference recording and browsing. Assistance to fact-finding has been evaluated using the Browser Evaluation Test, and a set of reference scores are now available. The analysis of these findings departs from the view that a system's lifecycle forms a closed loop, alternating requirements elicitation with design, implementation, and evaluation. Instead, the paper proposes a helix model that moves forward towards more and more refined systems, sometimes branching out to a new task.

1 Introduction

The design of technology for recording, processing, and browsing human meetings has become a significant research field in the past decade. Developments in this field have mainly been driven by the possibility of applying advanced signal processing methods, including multimodal processing, to data that is acquired using off-the-shelf capture devices. Large databases of meeting recordings have thus been gathered, with the potential for many more in the upcoming years. Front-end interfaces to such databases have been built thanks to multimedia retrieval algorithms and HCI technology. However, the field has often put applications before methodology, and therefore the definition of common tasks and benchmark data has lagged behind the design of particular systems.

This paper overviews the achievements of two long-term, multidisciplinary consortia[1] to demonstrate that a specific task, namely *assistance to fact-finding*, has become of significant importance to the field of meeting browsing. The specification of this task, which is relevant to users and developers of technology alike, has emerged gradually from a series of back-and-forth exchanges between representatives of these two groups, which will be discussed as follows.

2 Finding facts in meeting recordings

The paper will first review user studies related to meeting browsing that were conducted at several moments, with various populations and sets of questions. These studies show that user requirements alone cannot not lead from the start to a precise specification, as they depend strongly on the assumptions of each study – a chicken-or-egg problem. To become implementable, the user requirements had to be gradually focused towards the most promising task, which appeared to be fact finding or fact verification in multimedia meeting recordings.

Fact-finding answers significant user needs, and has sufficient generality to be of theoretical interest to researchers in the meeting processing area. The task is within reach of current multimedia processing techniques, as shown by the overview of the large range of meeting browsers that have been designed. In the course of mutual adjustment between user requirements and software development, a new application with end-user products has emerged: the acquisition, storage, and browsing of conference presentations.

Assistance to fact-finding can be quantitatively evaluated thanks to benchmark methods such as the Browser Evaluation Test (BET). As a result, a

[1]The Swiss National Center of Competence in Research IM2 – Interactive Multimodal Information Management (2002–2013), and the European AMI Consortium – Augmented Multiparty Interaction (2004–2010), both headed by the Idiap Research Institute.

set of reference scores has become available for comparison purposes, from several meeting browsers that were tested by human subjects, as well as from a fully automatic fact verification system.

To conclude, the analysis of these findings will depart from the view that user requirements, their implementation, and the evaluation of the resulting system form a closed loop. We submit that the articulation of these stages is better represented as a helix, which moves forward by alternating the requirements elicitation stage and the design / implementation / evaluation ones, towards more and more refined systems, sometimes branching out into a new helix that involves different tasks and systems.

3 Current practices for meeting archival

Studies of user needs are comparatively less frequent than specific proposals for meeting support tools, although capturing user needs normally initiates the development cycle of a software product. Two strategies have essentially been used to obtain specifications from user input. The first one has been to analyze the use of current information technology for meeting support, and to infer the needs that new technology could fulfill. The second one has been to ask users to describe functionalities that, if available to them in the future, would likely support their involvement in meetings better than existing technology does. This requires some guidance of the polled users, outlining more or less specifically the range of functionalities that can be expected from future technology.

Both approaches have advantages and shortcomings, as the examples below will show. The first approach leads to firm, verifiable conclusions regarding current practices, but inferring from them exact specifications for future tools, on the ground that they answer limitations of current ones, may require a considerable leap of faith. The second approach is more deterministic in turning expressed user needs into precise specifications, but is still faced with a dilemma regarding generality. On the one hand, if users are left free to imagine potential functionalities, then it might be difficult to agree on a prioritized list, and many suggestions might be far from implementation capabilities. On the other hand, if users are too constrained by feasibility issues (sometimes shown as a partly implemented architecture), then their answers might not reflect the most urgent needs, and the resulting software will merely reflect the designers' intuitions, with the risk of low utility or acceptance rates. Moreover, the results of the user studies are likely, in this latter case, to mix the evaluation of the existing specifications with the elicitation of new ones.

Two ethnographic studies of practices regarding the use of information from meetings, in a corporate context with series of project-related meetings, led to somewhat different conclusions – likely due to the different perspec-

tives of the experimenters [Jaimes et al., 2004, Whittaker et al., 2008]. Both studies interviewed a dozen people over several weeks or months, and the first study made a survey of 500 people, in order to explore the types of records and cues that people use to recall information from past meetings. The first study intended to explore the potential utility of visual information, while the second one focussed on more traditional records such as written minutes or personal notes, possibly based on transcripts of audio recordings. Therefore, while both studies confirmed the importance of structured meeting minutes for recall, they differed in many other conclusions.

In the first study, the polled users believed in the utility of audio-visual recordings for verifying or better understanding points in a meeting, and as an accurate overall record, while in the second one the users emphasized the limitations of official minutes for recalling specific details – a limitation somewhat overcome by private notes. Searching verbatim meeting records was a potentially challenging task: the first study showed that visual cues related to the meeting room and the participants facilitated recall, as did the list of topics discussed in the meeting, while the second study put forward the difficulty to retrieve important items such as assigned tasks or decisions, and emphasized the need for summaries rather than for full records.

Two other ethnographic studies [Cremers et al., 2007, Bertini and Lalanne, 2007] with respectively 10 and 100 users confirmed the previous insights. In order to retrieve information about a past meeting they attended, people use minutes and personal notes, though almost just as often they rely on personal recollection or even on emails and their attachments. The utility of audio-visual recordings alone was considered to be quite low, the main reason – for about half of the participants to the second study – being the time that is needed to go through the recording of an entire meeting. Given this constraint, it is of no surprise that recordings were viewed as useful to check what someone has said, especially in case of doubt, or as a proxy for people who missed an important meeting. Among the reasons why someone would need to review a past meeting, the most frequent ones are the need to prepare an upcoming meeting (which emphasizes an important property: the seriality of meetings), and the need to remember past topics, assigned tasks, or the date of the next meeting.

4 User requirements

One of the important conclusions that can be drawn from the four studies cited above is that raw audio-visual recordings of meetings appear to be of little use without tools that offer finer-grained access than current media players do. Starting from this observation, other studies have asked participants to imagine that they are using an "intelligent" search and navigation tool, and to describe the tasks that it could perform, or queries that

they would address to it. The variation of what subjects can be induced to imagine is illustrated in the studies cited below and summarized in Table 1. The studies include tasks that users could perform with the help of a system, tasks that a system could be expected to perform upon receiving a command, or formal queries over meeting data.

One of the above studies [Cremers et al., 2007] included a section in which eight users were asked to imagine an application generating smart meeting minutes from recordings. The most demanded pieces of information to include in such minutes appeared to be the arguments for decisions, the main topics and things to do, but also simply the meeting agenda and the names of the participants. When trying specifically to catch up on a missed meetings, users emphasized the need for a summary or gist of the meeting, together with a list of things to do, accompanied by a browser adapted to the smart minutes. In a query set with about 60 items collected from professionals [Banerjee et al., 2005], the most frequently requested item was the list of topics or themes discussed at a meeting.

Three large sets of queries were collected by members of our consortia [Lalanne and Sire, 2003, Lisowska, 2003, Wellner et al., 2005b] in various experimental settings. While the first collection included only developers of meeting technology and did not specify use cases for accessing recordings, the second one featured a mix of 28 participants, half of which had not been previously exposed to meeting technology. The participants could choose between four use cases – a manager tracking employee performance (5 subjects) or project progress (4), an employee missing one project meeting (12) or joining an ongoing project (7) – and were asked to state in their own words the questions that they would formulate to access a meeting archive. About 300 queries were collected and analyzed [Lisowska et al., 2004], with the purpose of inferring requirements for meeting processing, regardless of feasibility, such as the extraction of specific features from meeting media.

The main findings of this study concerned the type of information that users would look for: (1) queries related to the interaction between participants, bearing on elements such as decisions, questions, discussions, or disagreement; and (2) queries bearing on items that are conceptually part of meeting activities, such as dates, people, documents, presentations, and including also global and local discussion topics. These categories, and their sub-divisions, appeared to be overlapping by necessity, as queries can target the communicative and the content dimensions of a meeting fragment or utterance at the same time. Answering the queries requires topic detection, e.g. terms or significant keywords, named entity recognition, but also an understanding of the interaction structure, e.g. in terms of speech acts or decision processes, which in many cases far exceeds current processing capabilities. A sizeable number of queries were directed towards elementary meeting items, such as presentations, agenda and dates, and can be answered using simple processing of meeting recordings.

A different perspective on query analysis focuses on requirements for understanding queries, e.g. by converting them into a formal expressions that can be processed by a computer. Language-based queries have not been investigated in great detail, most studies assuming that browsing interfaces could assist the user in formulating a complex formal query without the need for analyzing linguistic input. Nevertheless, a large-scale Wizard-of-Oz study with 91 subjects [Lisowska et al., 2007], using a partially-implemented interface, has made a number of observations regarding the modalities most often used to access the archive, to complete tasks assigned by experimenters. The study showed that exposure and training had a strong impact on the way people used modalities to formulate queries – speech, written language, or mouse clicks – with no "natural" combination standing out. Speech was slightly preferred over other modalities to interact with the system, as the system appeared to understand it acceptably, thanks to a dedicated human "wizard" in the background.

Query analysis can be done also on the data obtained using the BET query collection procedure, described below [Wellner et al., 2005b, Popescu-Belis et al., 2008a]. Subjects were asked to formulate "observations of interest" regarding three recorded meetings from the AMI Corpus [Carletta, 2007]. The observations had to capture aspects that the subjects thought to have been "important to participants", which are potential targets for subsequent search. Users were explicitly asked to mark observations as either local or global, i.e. for a given moment, a short interval ,or throughout the meeting. However, the design of the collection procedure using an audio-visual meeting player encouraged observers to formulate many more local than global queries – a fact that might not be representative of the natural proportion. In the non-consolidated set of 572 statements from 21 observers, 63% of the statements refered to specific moments, 30% to short intervals, and only 7% were about the entire meeting. As for the content, five classes can be distinguished: statements about decisions (8%), about facts stated by participants (76%, including arguments leading to decisions), and about the interaction process or the media used by participants (11%); additionally, statements about the agenda and about the date of the following meeting were infrequent (2% each) but mentioned by most subjects. If the same analysis is made over the 251 statements mentioned by at least three subjects each, then the proportions of statements regarding decisions, agenda and dates increase to 13%, 4% and 3% respectively, while those related to process/media decrease to 2% and those regarding facts or arguments remains constant.

Author	NS	Method	Focus	Summary of findings
Jaimes et al. [2004]	15	interviews	practice	Utility of audio-visual recordings for verifying or better understanding points in a meeting.
	519	questionnaires	practice	Importance of visual cues for recall.
Whittaker et al. [2008]	12	interviews	practice	Importance of private notes, and need for summaries.
Bertini and Lalanne [2007]	118	questionnaires	practice / needs	Utility of raw audio-visual recordings is quite low, except for persons who missed a meeting or to find back specific information.
Banerjee et al. [2005]	12	interviews	practice	Thematic content is an important aspect of user queries.
Cremers et al. [2007]	8	interviews	needs	Need for a summary and list of things to do.
Lisowska [2003]	28	elicitation of queries	needs	Training has a strong influence on the strategies and modalities employed by users to review meetings.
Wellner et al. [2005b]	21	elicitation of observations of interest	needs	The browser used to collect observations can bias the collection. In experiment, most observations were local, and arguments leading to decisions.

Table 1: Comparison of user studies for meeting retrieval technology ('NS' stands for number of subjects).

5 Towards a specification for meeting browsing

This review of user studies, summarized in Table 1, shows that requirements for meeting archival and browsing technology are multi-faceted, but that their main dimensions are now well understood. Requirements can be categorized in terms of the targeted timespan (utterance, fragment, meeting, series), targeted media (audio, video, documents, presentations, emails), complexity of information that is searched for (present in the media vs. inferred), but also in terms of the complexity and modalities of the queries.

Two main categories of applications answer part of these requirements each: meeting summarization and meeting browsing. Meeting summarization or gisting has appeared to be a challenging but tractable task, using mainly audio recordings, and several meeting summarization systems [Zechner, 2002, Murray et al., 2005] and evaluation techniques [Murray et al., 2008] have been proposed. However, the main focus of our consortia has been on meeting browsers for fact finding or verification, as they answer the most frequently mentioned user needs, and raise important challenges for multimedia processing. Therefore, our user studies have been gradually narrowed down towards the elicitation of specific fact-finding tasks. However, these studies did not lead to a unique specification, as they depended greatly on how subjects were prompted to respond and how their answers were analyzed. Several meeting browsers, described below, have been implemented to answer user requirements, although they were never formally derived from them. In addition, benchmarking methods grounded in user studies were designed as well.

Meeting browsing as illustrated in Figure 1 appears thus to be a significant transversal application, striking a good balance between answering user needs, feasibility, and generality. Still, development could be made even more user-driven in the future, with a number of challenges to be addressed [Tucker and Whittaker, 2005, Section 3.4]. As user studies are notoriously difficult to generalize, a large number of studies are needed to circumscribe the range of options for meeting archival and access technology, and it is difficult to expect "definitive" user studies, all the more that the underlying technology evolves continuously. Of course, it is also likely that user studies carried out by private companies for the development of proprietary products are never published, as they offer companies competitive advantage, but it is also possible that many products, including very successful ones, are actually developed without a clear view of the user needs that must be answered – in this respect, reference to published studies should be beneficial.

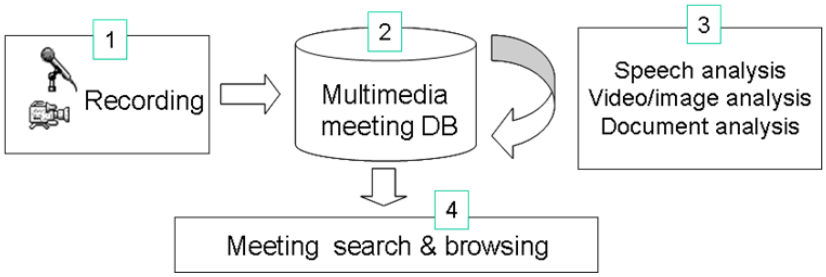

Figure 1: Generic architecture of a meeting processing and retrieval system.

6 Meeting browsers

Two main types of applications partially answer the above requirements. On the one hand, meeting summarization systems offer an abstracted view of a meeting, structured for instance around its main topics – as in the early 'Meeting Browser' from the Carnegie Mellon University [Waibel et al., 1998] – or around the tasks or 'action items' that were assigned – as in the CALO browser [Tür et al., 2010] (Cognitive Assistant that Learns and Organizes, US project, 2003–2008). On the other hand, other meeting browsers are intended to help users with fact finding or verification – e.g. to check figures, assigned tasks, decisions, or document fragments – although they can also be used to sample a meeting for abstractive purposes. Recent surveys [Tucker and Whittaker, 2005, Bouamrane and Luz, 2007, Yu and Nakamura, 2010] include examples of both types, which are also classified according to the main rendered modality, as Tucker and Whittaker [Tucker and Whittaker, 2005] do, or according to basic or enhanced functionalities, as Yu and Nakamura do[Yu and Nakamura, 2010].

A large number of meeting browsers supporting fact-finding in audio-video meeting recordings have been designed, both within [Lalanne et al., 2005] and outside [Tucker and Whittaker, 2005, Whittaker et al., 2008] our consortia. Such browsers locate specific bits of information within a meeting that typically lasts 30 to 60 minutes, based on a variety of features extracted from multimedia recordings, e.g., speech transcript, turn taking, documents in verbal focus, slide changes, or handwritten notes. The browsers can be classified according to the major modality that is used as an index for locating relevant excerpts from a meeting and triggering their playback. Speech-centric browsers take advantage of the audio recordings and/or their transcript, possibly accompanied by higher-level annotations such as named entities, thematic episodes, or dialogue acts. Conversely, document-centric browsers take advantage of document content, recognized through various analysis methods, and annotations such as slide changes.

The range of meeting browsers developed within our consortia [Lalanne

et al., 2005, AMI, 2006] answered partly the user requirements found above, but also followed the developers' views of what functionalities could be useful given the technical components that were available for integration and testing at various stages of each project. The design of such browsers requires in fact the availability of complete transversal systems that access and analyze meeting data, in order to build high-level indexes that are used by interactive multimodal user interfaces. Although it is possible to reuse the components of such a transversal system with various front-end meeting browsers, in many cases each browser is accompanied by its own system. Reusability is more frequent at the level of data and its reference annotations: the AMI Corpus [Carletta, 2007] and its annotations in more than a dozen dimensions is used by many of the browsers described below.

The description below makes a graded progression between partially implemented browsers, through fully implemented ones but which work over human-processed meetings (e.g. using reference annotations from the AMI Corpus), to fully automatic ones which do not require human intervention at all (apart from organizing the recorded data). While partially implemented browsers serve to capture additional user needs, the fully automated ones can be submitted to quantitative evaluation. Table 2 summarizes the main features of the browsers developed within our consortia, while the most representative ones are shown in Figures 2 and 4. Each browser renders a different subset of media/modalities of the meeting recordings, and offers different criteria when searching meetings for specific facts. To support the development of these browsers, two toolkits were created: JFerret [Wellner et al., 2005a, 2004, 2005b] allows the definition of reusable graphical components that access meeting annotations, while HephaisTK [Dumas et al., 2009b] enables the rapid prototyping of multimodal interfaces, including a multimodal fusion engine with state-of-the-art performance [Lalanne et al., 2009, Dumas et al., 2009a].

6.1 Speech-centric browsers

Some of the simplest browsers implemented in the JFerret framework are two audio-based browsers [AMI, 2006] which provide access to audio recordings, with speaker segmentation and slides, and enhance speech browsing in two ways: the *Speedup browser* accelerates audio playback while keeping speech understandable by avoiding the chipmunk effect; and the *Overlap browser* plays two different parts of a meeting in the left vs. right channels, assuming that the user will take advantage of the cocktail party effect to locate the most relevant channel and then adjust the audio balance to extract the interesting facts. The *JFerret browser* [Wellner et al., 2005a,b], in fact a sample implementation illustrating the main capabilities of the JFerret framework, offers access to audio, video, slides, ASR transcript, and speaker segmentation – as exemplified in Figure 3. The browser has been

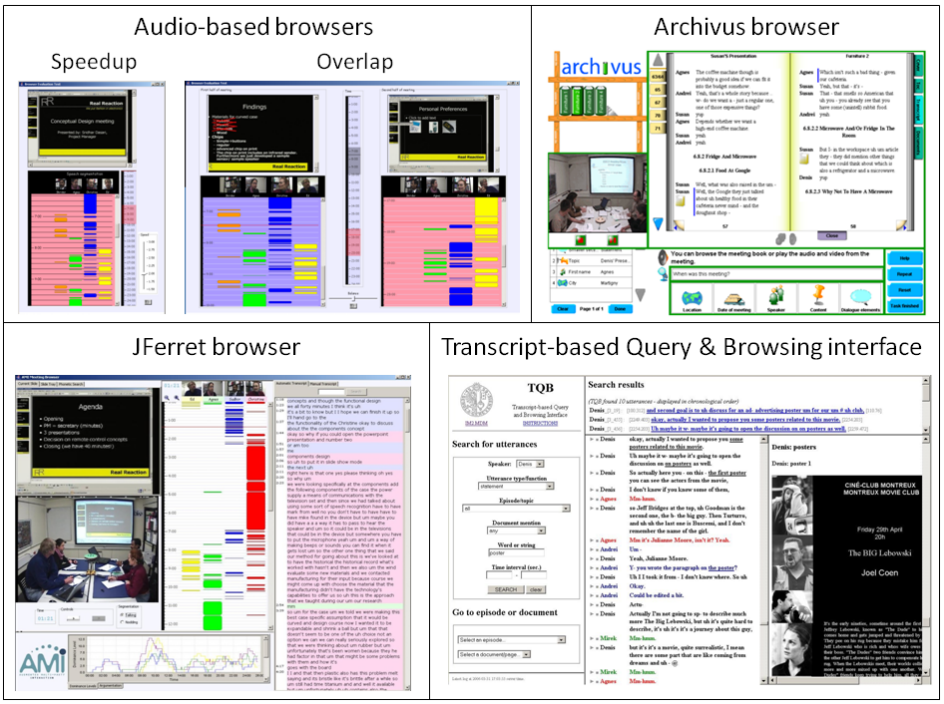

Figure 2: Speech-centric meeting browsers from our consortia: clockwise from top left, audio-based browsers [Popescu-Belis et al., 2008a], Archivus [Lisowska et al., 2007], TQB [Popescu-Belis et al., 2008a], and JFerret[Wellner et al., 2005b]. A larger view of JFerret is provided as an example in Figure 3.

extensively demonstrated and studied by different teams, e.g. [Whittaker et al., 2008, Section 5]. The *Transcript-based Query and Browsing (TQB) interface* [Popescu-Belis and Georgescul, 2006, Popescu-Belis et al., 2008a] is another speech-centric browser, which uses a number of manual (reference) annotations in order to compare their respective utility to users: manual transcript, dialogue acts, topic labels, and references to documents.

Archivus [Ailomaa et al., 2006, Melichar, 2008] is a meeting browser based that enables multimodal human-computer dialogue, thanks to a Wizard-of-Oz approach which allows for partial implementation only, in order to gather user requirements [Lisowska et al., 2007] as mentioned above, especially in terms of modality use. Archivus also uses reference transcripts enriched with annotations (speaker segmentation, topic labels, documents) to answer user queries that are expressed as a set of attribute/value constraints over one or several meetings. An implementation using a standalone dialogue engine with a multilingual front-end and a touch-screen on a mobile device was also achieved for a subset of the Archivus search attributes,

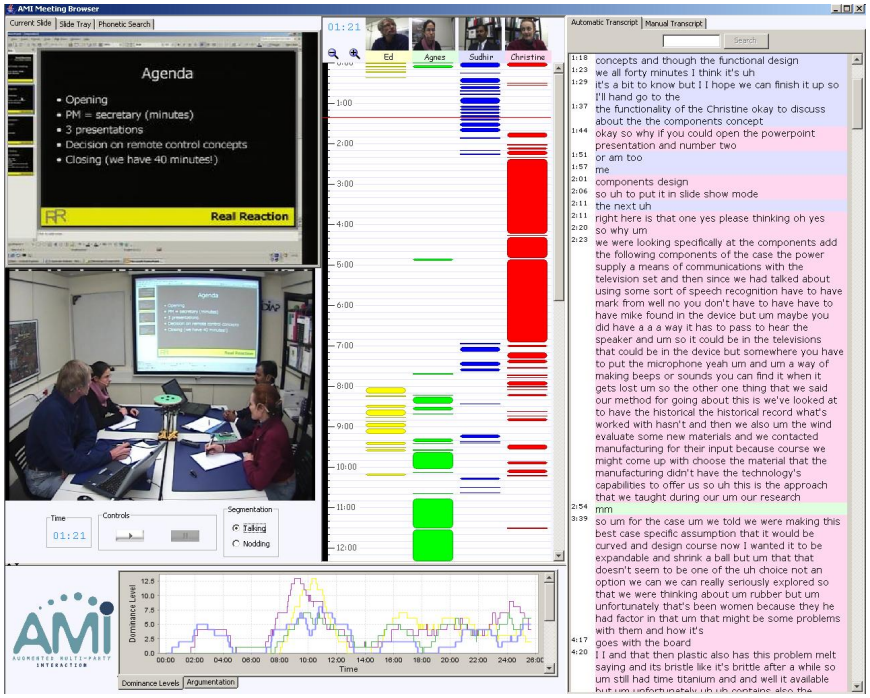

Figure 3: A meeting browser prototype built using the JFerret framework [Wellner et al., 2005b]. Clockwise from top left: slide display, speaker turns, transcript (automatic or, for testing, manual), dominance levels computed from multi-modal features, video of room. All time-dependent components are synchronized to the main timeline displayed with the speaker turns.

as the *Multilingual, Multimodal Meeting Calendar (M3C)* [Tsourakis et al., 2008].

6.2 Document-centric browsers

FriDoc [Lalanne and Ingold, 2004], followed by *JFriDoc* [Rigamonti et al., 2006], are document-centric browsers that exploit the alignments between printed documents and speech. They contain the documents discussed during the meeting, dialogue transcripts, slides and audio-video streams. In these browsers, clicking on a document section places the audio/video sequences at the moment when the content of this document block is being discussed, and reversely. Similarly, the *ViCoDe* prototype (for *Video Content Description and Exploration*) focuses on video similarity between sentences and on relevance feedback to propose a novel manner for browsing meetings [Marchand-Maillet and Bruno, 2005]. The *FaericWorld system* [Rigamonti et al., 2007] enhances the document-based browsing strategy with cross-meeting representations of documents and links. For each collection

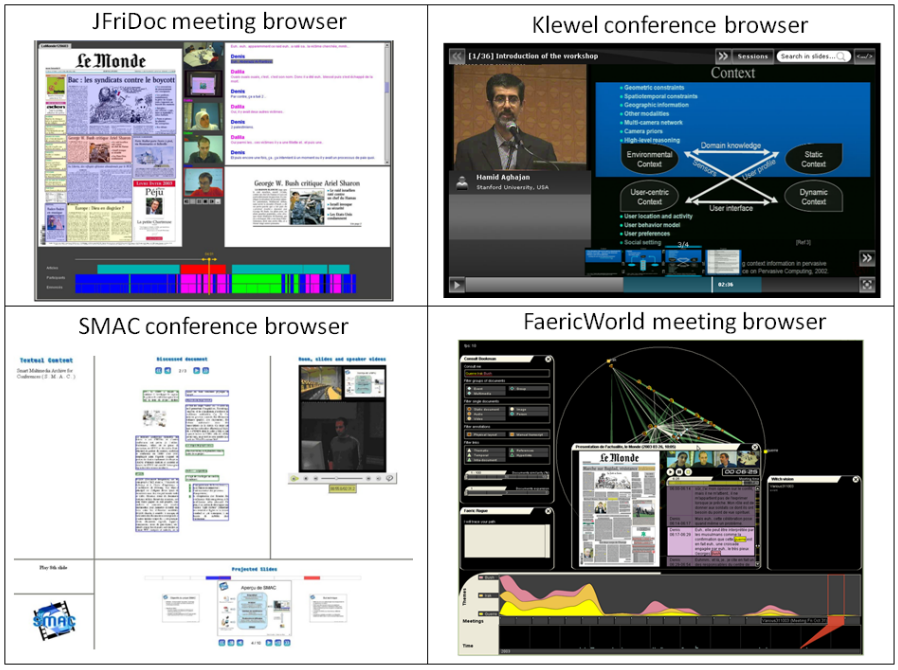

Figure 4: Document-centric meeting browsers and conference browsers from our consortia: clockwise from top left, JFriDoc [Rigamonti et al., 2006], Klewel, FaericWorld [Rigamonti et al., 2007], and SMAC.

of meetings, links between all the categories of multimedia documents belonging to the meetings are automatically calculated. Users can then query the system with full text search or directly browse through links, thanks to interactive visualizations. Further, the *WotanEye system* [Évequoz and Lalanne, 2009] has been developed to enable ego-centric access to meeting fragments using personal cues, such as the user's social network.

6.3 Automatic Browsers

Automation of meeting browsers was studied along two axes. The first one is to provide access to past meetings in a query-free manner during an ongoing meeting, by using a just-in-time information retrieval approach, in the *Automatic Content Linking Device (ACLD)* [Popescu-Belis et al., 2008b, 2010]. The words currently spoken are recognized by a real-time ASR system [Garner et al., 2009] and serve to launch queries to a database of past meeting snippets, indexed by their words – also recognized by ASR. The meeting snippets thus retrieved can be viewed with any suitable device, e.g. with one or more of the rendering components of JFerret. The second axis aims at automated question answering over meetings, and has been explored through the *AutoBET browser* [Le and Popescu-Belis, 2009], which

attempts to discriminate the true vs. false statements about a meeting from a BET pair (see below). The system first identifies the passage of the meeting transcript that is most likely to contain the discriminant information, then hypothesizes which statement is true based on lexical matching between statements and the passage. The scores of AutoBET, along with those of several other browsers used as fact-finding assistants, will be discussed below.

	Speedup	Overlap	JFriDoc	Archivus	TQB	JFerret	FaericWorld	Klewel	SMAC
Access to media									
Audio	x	x	x	x	x	x	x	x	x
Video			x		x	x	x	x	x
Slides	x	x	x	x	x	x	x	x	x
Documents				x	x	x	x		x
Input features									
Manual transcript			x	x	x	x	x		
Automatic transcript			x						
Speaker segmentation	x	x	x	x	x	x	x		x
Slide content from OCR								x	x
Slide change	x	x	x			x	x	x	x
Manual dialogue acts				x					
Structure of documents						x	x		
Multiple meetings					x		x	x	x

Table 2: Comparison of meeting browsers according to the media accessed from recordings and the features extracted from them.

7 Conference recording and browsing

Despite the large number of research prototypes, there are no commercial meeting browsers that assist users with fact-finding beyond standard replay capabilities. This is all the more surprising since successful web conferencing hardware and software solutions are available, some of which with recording capabilities. This is possibly due to the divergences in user requirements discussed above, and to the advanced technology needed for recording and especially for processing meetings.

Therefore, in the course of the mutual adjustment between user requirements and software development, our family of meeting browsers has evolved towards two products that answer slightly different user needs. Namely, conference recording and browsing has emerged as a promising application to

consider, as a by-product of meeting indexing and browsing, though the first products were as much based on the designers' objectives as they were on unbiased users' needs.

Conference recording and browsing devices appeared to be easier to deploy at the production level, as they imply fewer capture devices to synchronize than in meeting rooms, and a comparatively smaller amount of data to store and analyze. A number of robust indexes can be extracted, such as slide changes, text from slides, or slide/audio/video synchronization, which are already helpful for browsing, and can provide support for fact-finding.

Two systems for conference recording and browsing have reached production stage, using several technologies developed in the consortium. One system is commercialized by an Idiap spin-off company (Klewel),[2] and the other one was developed by the University and the Engineering School in Fribourg and the CERN in Geneva, within the SMAC project (Smart Multimedia Archive for Conferences), and is in production mode at these institutions. Both systems capture, index and render audio, video and slides using off-the-shelf capture devices and web-based interfaces. The first one is currently enhancing its indexing capabilities using ASR to recognize the speaker's words, while the second one uses document analysis and alignment techniques [Lalanne et al., 2004] to automatically hyperlink, in scientific talks, the fragments of the scientific article that is being presented with the related audio-video sequence.

8 Evaluation of meeting browsers

8.1 Evaluation frameworks and campaigns

The evaluation of meeting browsers, as pieces of interactive software, is in principle related to a precise view of the specifications they answer, as it appears for instance from ISO's[3] definition of quality, as the extent to which a system fulfils stated or implied user needs [ISO/IEC, 2001, Azuma, 2001]. However, as it should have become clear by now, many meeting browsers only answer quite generic requirements, and include some components that originate mainly in their developers' choices. It is also difficult to define specific functionalities that all browsers should implement, in order to perform internal or external evaluation (in ISO terms). Evaluation efforts have thus been concentrated on task-based approaches, close to evaluation in use [Bevan, 2001], though not in end-user environments. The challenges were related to the definition of tasks that could be considered by the community

[2]The Klewel/Idiap PAS was adopted by the ACM Digital Media Capture Committee, following the successful recording and distribution of the CHI 2007 conference. The company received the European Seal of e-Excellence, from the European Multimedia Forum, at CeBIT 2008.

[3]The International Organization for Standardization.

to be enough representative of the meeting browsing activity. One can distinguish, as above, tasks related to meeting abstraction and tasks related to fact-finding; in addition, a task can either involve a fully autonomous tool, or, more frequently, human subjects operating an interactive assistant.

The evaluation of interactive software is known to be a challenging endeavor, in particular for multi-modal dialogue systems [Gibbon et al., 2000, Dybkjær et al., 2004]. The main quality aspects that are evaluated in a task-based approach are *effectiveness*, i.e. the extent to which the software helps the user to fully complete a task, *efficiency*, related in particular to the speed with which the task is completed, and *user satisfaction*, which is measured using questionnaires.[4]

The evaluation of the fact finding functionality of meeting browsers has drawn inspiration from the evaluation of question answering (QA) systems, which has provided quantitative results in a framework that is easier to setup and reproduce than generic "meeting improvement" approaches. QA systems have been evaluated starting with the TREC-8 (1999) QA track [Voorhees and Tice, 1999, Voorhees, 2001], using a set of questions with known answers, and simply measuring how many answers provided by a system matched the desired ones. The challenge in this approach is to obtain non-biased questions; in the 1999 QA track, 1,337 questions were obtained from multiple sources (participants, assessors, organizers, and one Web-based QA system) from which 200 were selected for the campaign. At TREC 2003, the test set of questions contained 413 questions of three types (factoid, list, definition), which had been drawn from AOL and MSN Search logs [Voorhees, 2003]. Similarly, multilingual QA has been a track of the Cross-Language Evaluation Forum (CLEF) since 2003, and despite variations, the factoid QA task has been continuously tested, e.g. in 2009 in the field of European Legislation (ResPubliQA [Penas et al., 2009]); the monolingual aspect (same language for question and documents) has however been prevailing in each campaign.

Interactive QA systems were evaluated in another CLEF track, iCLEF [Gonzalo et al., 2006], as well as in the ciQA task of TREC proposed in 2006 and 2007 [Dang et al., 2007]. Overall, systems-plus-humans were evaluated for accuracy over a large set of questions defined by the experimenters, and differences in performance were used to infer a ranking of the systems. The iCLEF task followed this approach in 2001–2005, although very few participants could afford the cost of setting up such user-centric experiments (only three systems participated in 2005, on a 16-question task), then moved away from quantitative evaluation and towards log analysis of an image retrieval task. In the TREC ciQA task, the assessors interacted with online systems

[4]The PARADISE approach to dialogue system evaluation [Walker et al., 1997] has shown that user satisfaction stems from task completion success and from dialogue cost, therefore it makes sense to focus on effectiveness and efficiency as means to achieve user satisfaction.

for five minutes per question, two stages of each system being compared in a two-week interval.

8.2 Evaluation within the AMI Consortium

The most generic approach to meeting browser evaluation considers such browsers simply as tools that "improve meetings", which can be used between meetings (or even during meetings) to retrieve previously stated information. The improvement of meetings can be measured through a number of parameters related either to their outcome (e.g. an optimal decision or not), or to the process itself (e.g. a pleasant atmosphere). Both types of indicators have been shown to be statistically reliable for collocated meetings [Post et al., 2008] and in a test bed with one remote participant [Post and Lincoln, 2008]. This approach has been used in a large experiment [Post et al., 2007, AMI, 2006] with 22 teams of four subjects, holding series of four meetings each, which were aimed at the design of a remote control as in the AMI Corpus scenario [Carletta et al., 2006, Carletta, 2007]. The experiment compared meeting browsers in four conditions: (1) no browser at all; (2) JFerret browser with manual transcripts; (3) same as (2), but with an automatically generated abstract; and (4) task-based project browser with access to recordings. The results showed that the third condition lead to the highest perceived quality in terms of meeting success.

Turning now to more specific functionalities, evaluations of meeting summarization have been carried out using either reference-based or task-based approaches. In one of the evaluation experiments carried out by Whittaker et al. [Whittaker et al., 2008], the utility of compressed speech for meeting summarization by humans was tested experimentally. Subjects were asked to rank the importance of utterances from a meeting recording that they heard in various compresed formats, accompanied or not by a player; the ranking of utterances, in comparison to a gold-standard extractive summary, provided information on the utility of each format or player. Another approach by Murray et al. [Murray et al., 2008, 2009] used a "decision audit task" in which subjects used five different types of summaries to analyze how a particular decision was arrived at in a series of meetings. The analyses were judged by another series of subjects, thus providing a set of final scores, along with log analysis and post-task questionnaires, which were informative about the quality of each initial summary. In particular, the study demonstrated that automatic summaries done with ASR transcripts, are still useful to decision analysis, though less than human abstracts.

8.3 Current approaches to the evaluation of meeting support technology

By many accounts [Bouamrane and Luz, 2007, Yu and Nakamura, 2010], the evaluation of meeting support technology is a challenging task, though it is unavoidable to demonstrate appropriateness of design, or to compare several designs, interaction paradigms, or meeting analysis tools. As synthesized by Yu and Nakamura [Yu and Nakamura, 2010, p.11–12], "the criteria used to evaluate a smart meeting system include user acceptance, accuracy [of recognition mechanisms], and efficiency [. . . i.e.] whether a browser is useful for understanding the meeting content quickly and correctly." While accuracy of recognition is not by itself a measure of browser quality (though it influences the browsing experience), the two other criteria reflect two opposing, though not incompatible, views of evaluation. Indeed, as formulated by Abowd et al. [Abowd et al., 2002, p.56], "it is not clear that [performance and efficiency] measures can apply universally across activities", and one must "consider how to undertake assessment that broadens from existing task-oriented approaches."

Several studies of individual meeting browsers have considered both approaches to evaluation. The Filochat system of the early 1990s [Whittaker et al., 1994] was one of the first browsers for time-aligned speech recordings and personal notes; a user study demonstrated the usability of the system and helped to assess desirable and undesirable features, while laboratory tests compared three conditions (notes only, record only, or Filochat) by measuring accuracy and speed of subjects answering factual questions about what they heard. In the Xerox PARC system for "salvaging" fragments of recordings in order to build accurate minutes [Moran et al., 1997], evaluation is based on observations of use over one year, demonstrating "how practices develop and differentiate" and how the system influences its users. The relation between accuracy of processing and user behavior has also been mentioned for the CALO action item detector and browser (see [Tür et al., 2010] and references therein). Finally, for the AMI/IM2 JFerret meeting browser (here augmented with automatically generated abstracts), a large experiment [Post et al., 2007] with 27 teams of four people holding series of meetings has shown that it outperformed two other browsers (and no browser at all) in terms of impact on meeting success.

The needs for comparing meeting browsers (at the same period or over time) are however better satisfied by efficiency-oriented evaluations, which provide a more controlled setting and a standardized, easier to apply protocol than user studies do. Efficiency can be measured over benchmark tasks that are representative of the meeting browsing activity, in realistic contexts.

8.4 The Browser Evaluation Test (BET)

Within our consortia, the Browser Evaluation Test (BET) framework was designed and applied to collect and use meeting-related questions for browser evaluation [Wellner et al., 2005b]. Unlike QA evaluation, BET questions are pairs of parallel true/false statements which are constructed by neutral human observers, who (1) view a meeting recording, (2) write down observations of interests about it, i.e. the most salient facts, and (3) create for each statement a plausible but false counterpart. This procedure was adopted to avoid biasing queries with developers' interests for specific browsing functionalities. Therefore, observers were always external to the consortium, and were asked to select what they thought was important to the participants in the viewed meeting recordings, and indicate whether this was a local or global piece of information. Pairs of statements referring to the same piece of information were consolidated into groups by experimenters, who picked one representative per group; an importance score was automatically computed from the observers' rating and the size of each group.

Three meetings from the AMI Corpus [Carletta et al., 2006, Carletta, 2007] were selected for the BET observation collection procedure, resulting in about 570 raw observations from around seven observers per meeting, and a total of 350 final pairs of true/false statements. The average size of the consolidated groups is around two statements per group, i.e. each statement was mentioned by about two observers; in practice, however, when considering only the statements that human subjects had the time to process in the experiments below, each statement was mentioned on average by five observers. Examples of most frequently mentioned pairs of true/false observations are given in Table 3.

BET pairs can be used to evaluate browsers used by humans to distinguish true from false statements, but can also serve to evaluate automatic browsers designed to perform this distinction. In the first case (by far the most frequent ones as it does not impose specific constraints on the browsers), subjects discriminate the BET pairs in sequence, in principle by order of decreasing intrinsic importance. This order was verified to contain no hidden dependencies (earlier statements disclosing the answers to future ones), and ensures that, if time is limited, the most important facts of the meeting are searched for.

Apart from observing the subjects' behavior with the browser, and measuring their satisfaction using post-experiment questionnaires, two main scores can be computed. *Precision* is the number of correctly discriminated pairs, and indicates effectiveness, while *speed* is the number of pairs processed per unit of time, and indicates efficiency.[5]

[5]The average speed is typically not just the arithmetic average of sveral speed values – because time is an additive quantity, but not speed – but should be calculated from the average time to answer a question.

Meeting	**N**	**T/F**	**Statement**
Movie club to discuss the next movie to show	1	true	The group decided to show *The Big Lebowski*.
		false	The group decided to show *Saving Private Ryan*.
	2	true	Date of next meeting confirmed as *May* 3^{rd}.
		false	Date of next meeting confirmed as *May* 5^{th}.
Technical meeting to design a remote control	1	true	According to the manufacturers, the casing has to be made out of *wood*.
		false	According to the manufacturers, the casing has to be made out of *rubber*.
	2	true	*Christine* is considering cheaper manufacture in "other countries" [...]
		false	*Ed* is considering cheaper manufacture in "other countries" [...]
Lab meeting to furnish a new reading room	1	true	*Susan* says halogen light is very bad for reading in.
		false	*Agnes* says halogen light is very bad for reading in.
	2	true	The group decide they need at least *two* lamps.
		false	The group decide they need at least *four* lamps.

Table 3: Examples of BET pairs of statements for three meetings, with differences between true and false versions highlighted here for ease of understanding.

The acceptance of the BET as a valid test protocol must also acknowledge a number of possible biases or limitations. First, as any other evaluation method, the BET should check to what extent browsers conform to the user requirements presented above. In the case of the BET, the elicitation method biases these requirements towards fact finding or verification, as explained above, while other requirements elicitation study have emphasized higher-level elements of interest such as action items, topics or decisions, which are possibly under-represented in the current BET set, although a different set could be elicited with an inverse bias. Moreover, unlike many user-oriented evaluations (including those cited above), the BET observers and the BET subjects *are not* chosen among the participants to the meetings, although the observers are encouraged to make observations that would have been of interest to the participants. Therefore, as acknowledged above, these requirements and the related evaluation task are intended for "null-context" users, and cannot be used to compare directly a meeting browser with more subjective memorization devices such as personal notes taken during a meeting, although comparison with the use of third-party notes can be made. The somewhat focused spectrum of the BET is the price to pay in order to ensure reproducibility of the method, enabling comparison across browsers.

Browser	Condition	NS	T(s)	CI	P	CI
Audio-based	Speedup	12	99	26	0.78	0.06
browsers [AMI, 2006]	Overlap	15	88	23	0.73	0.08
JFerret	BET set (pilot)	10	100	*43*	0.68	*0.22*
[Wellner et al., 2005b]	5 gisting questions	5	<180	0	0.45	*0.34*
[Whittaker et al., 2008, p. 210]	5 factual questions	5	<180	0	0.76	*0.25*
TQB	1st meeting	28	228	129	0.80	0.09
[Popescu-Belis et al., 2008a]	2nd meeting	28	92	16	0.85	0.06
	Both meetings	28	160	66	0.82	0.06
FriDoc	With speech /	8	113	n/a	0.76	n/a
[Rigamonti et al., 2006]	document links					
	Without links	8	136	n/a	0.66	n/a
Archivus	T/F questions	80	127	*36*	0.87	*0.12*
[Melichar, 2008, Ch. 6.6]	Open questions	80	n/a	n/a	0.65	*0.22*
AutoBET	Movie club meeting	5fCV	<1	n/a	0.57	0.06
[Le and Popescu-Belis, 2009]	Remote control meeting	5fCV	<1	n/a	0.64	0.18

Table 4: Comparative results of several meeting browsers evaluated in similar conditions. NS is the number of subjects in each evaluation. T is the average time (in seconds) needed by subjects to answer a question, and P is the average precision, or proportion of correctly answered questions. Confidence intervals at 95% are absolute values; when they could not be determined, standard deviations are given instead (in italics). For AutoBET, confidence intervals are computed using five-fold cross validation ('5fCV').

8.5 Comparisons of BET scores

Several browsers have been evaluated using the BET questions, with more than 100 subjects passing the BET in various conditions, and the BET was confirmed as a good performance indicator for the fact finding task. The results obtained by the above-mentioned browsers are synthesized in Table 4 in terms of precision and speed, with 95% confidence intervals when available (or standard deviations), and the corresponding graphical representation is shown in Figure 5. The Table includes various conditions and states the number of subjects for each of them.

Comparison across scores must be taken with a grain a salt, given that baselines are not all at 50% (random binary choice), timing is variously constrained, and the subjects' compentencies and training differ across experiments; in addition, many browsers require some human preparation of the data. Therefore, the goal is not to point to the "the best browser", but to provide a well-founded overview of current state-of-the-art performance in meeting browsing for fact finding. The scores can be used for future comparison, with two reservations: (1) large differences in performance across

subjects, leading to large confidence intervals, and low statistical significance of differences in scores; (2) some experimental variability that prevents strict comparisons across conditions. Such comparisons, indeed, are licensed only if the same questions were used, in the same order, on comparable groups of subjects, trained in similar conditions, and having the same amount of time at their disposal.[6] These conditions are rarely met, except in strictly controlled evaluation campaigns, which have yet to be organized for meeting browsers.

The audio-based browsers were evaluated with a group of native English speakers at the U. of Sheffield using the standard BET [AMI, 2006]. JFerret browser was first evaluated as a pilot experiment with the standard BET pairs [Wellner et al., 2005b], and re-tested later, still a the U. of Sheffield, with five BET-inspired factual questions and five questions that required gisting [Whittaker et al., 2008, p. 210-211]; none of the conditions had a training phase. TQB was evaluated at the U. of Geneva [Popescu-Belis et al., 2008a] over the standard BET set, allowing 50% of meeting time, on two meetings; the training effect after one meeting could thus be measured. JFriDoc [Rigamonti et al., 2006] was also tested using BET-inspired questions at the U. of Fribourg. Subjects were allowed at most 3 minutes per question, and the experiment compared two conditions of JFriDoc, with vs. without speech / document links. Finally, the *Archivus* interactive multimodal browser was tested in a Wizard-of-Oz environment, with a very large number of subjects (from U. of Geneva and EPFL) answering 20 questions in 20 minutes, half true/false and half short-answer ones. The system's response time, included in the T value in Table 4, was on average 36 seconds, due to the wizards' latency in interpreting the user's actions before generating proper responses.

8.6 Results of BET evaluations

To synthesize, average discrimination time for a BET pair is around 2 minutes, with a 1.5–4 minute range: so, any significant improvement in the future should lower this limit.[7] Precision – generally against a 50% baseline except for open-answer conditions (JFerret and Archivus) – is in the 70–80% range, with higher values for browsers that make use of a lot of human-processed information (TQB and Archivus). More knowledge is thus helpful to increase precision, but this often means that subjects spend slightly more time to actually look for the right answer. The variability of human performance is higher for speed than for precision; in both cases, this variability challenges the statistical significance of comparisons.

These results can be compared with scores from interactive question

[6] Note that a group cannot be tested more than once over the same meeting.

[7] Sometimes, quick answers are from bored subjects who give up searching, so a method to detect this strategy in evaluation experiments should be found.

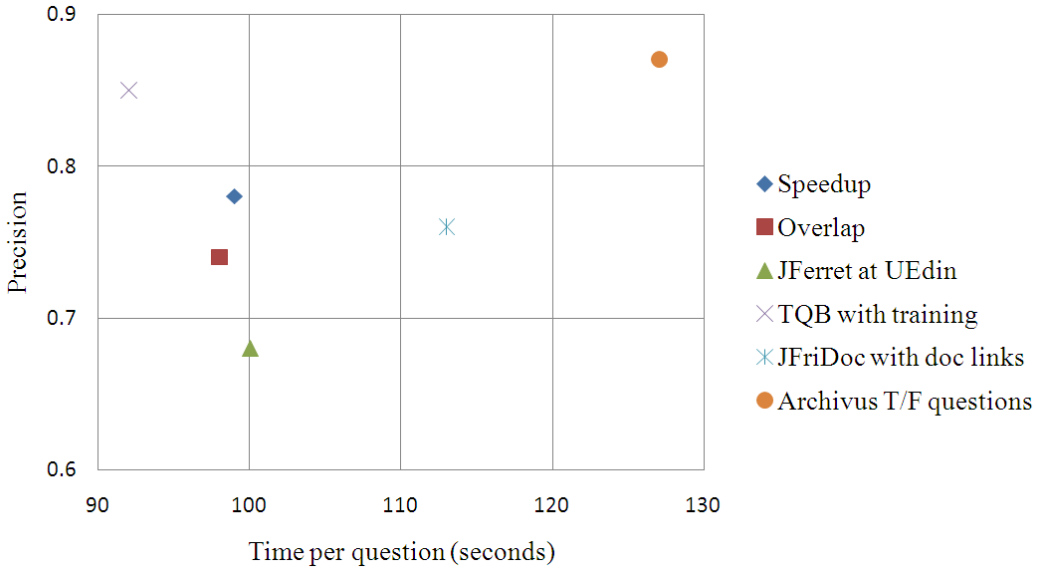

Figure 5: Representation of BET scores for the meeting browsers in Table 4.

answering campaigns, thus illustrating the difficulty to obtain reliable assessments of differences between browsers. For instance, three QA systems took part in iCLEF 2005 (the last edition to include comparative evaluation): eight users in each condition attempted to answer 16 open-ended factual questions in at most five minutes each. None of the six pairwise comparisons of conditions for the three systems were significant, because of the large variation across subjects. The system that submitted the most comparisons between conditions (four pairs) reached on average 55% accuracy (between 36% and 69% per condition), and 130 s speed (between 94 and 157 s).

The AutoBET fully automatic BET answering device [Le and Popescu-Belis, 2009] was also tested on the BET questions (last two lines of Table 4): while its speed was, as expected, far greater than any of the human subjects, precision remained well below human values, at 0.57±0.06 for one meeting and 0.64±0.18 for another one, only slightly above the 50% baseline. The identification of relevant passages is more accurate, at 0.55±0.14 and 0.62±0.16 compared to less than 0.01 by chance – but the system is penalized by the T/F discrimination module, which does not perform a fine-grained analysis of the found passage. To provide some reference points, passage identification reached 0.685 accuracy at TREC QA 2003 [Voorhees, 2003] (though of course on a much larger document set than here), and 0.68 at ResPubliQA 2009 [Penas et al., 2009], though most of the systems were well below the best scores (at TREC QA 2003, the average over 11 submissions was only 0.233). The best accuracy for precise answers to factoid questions

was 0.706 at TREC QA 2007 [Dang et al., 2007], but 8 out of 10 systems scored below 0.3, while at TREC QA 2003 the best accuracy was 0.700.

8.7 Lessons learned

The main lessons learned from the BET evaluations, apart from the reliability of the method, concern the available technologies that appear to be useful for meeting browsing. Transcripts are used intensively when they are of high quality, especially as users tend to perform keyword searches on them, thus pointing to the need for improved speech-to-text systems. However, other annotations of the transcript seem much less helpful.

The documents related to a meeting are relevant to fact finding, if available, especially when shown along the meeting's timeline, e.g. using automatic slide change detection and speech/document alignment. Slides can even compensate partly for the lack of transcript, as shown by audio only browsers, which score only slightly below transcript-based ones. The video recordings were the least helpful media for fact finding in our experiments.

Finally, learning effects appeared to be important: one training session improved the subjects' performance quickly, and conditioned their choice of modalities for browsing – which is good news for product designers, but poses some problems for the design of comparative evaluation experiments.

9 Synthesis: lifecycle of meeting browsers

The previous sections have outlined the main achievements in the definition and implementation of meeting browsers, which were carried out by a large number of teams, in two consortia, over eight years. The dialogue between teams in charge of requirements elicitation, of design and implementation, and of evaluation has been particularly active, but the global picture of the process has yet to be drawn. This section proposes a *helix model* of the resulting software development process, and outlines its main stages, keeping in mind that unlike commercial software products, meeting browsing software has been mainly developed for research purposes, in close relation to research on the analysis of multimodal human-human communication. Therefore, no specific customers were targeted for deployment, even though, in at least two cases, browsers have reached commercial stage for archives of recorded talks.

The experiments described above, from requirements elicitation for meeting support tools and in particular fact-finding assistants, along with the design and implementation of browsers, and their task-based evaluation, indicate that the engineering of a meeting browser is likely to be a complex software development process (SDP) [Sommerville, 2007, Chapter 2]. In the waterfall model of software engineering, users have the primary role of formulating the requirements for a task, which developers then attempt to satisfy

with the software product, evaluated against specifications [ISO/IEC, 2001]. However, collecting user requirements for a problem such as meeting browsing cannot lead directly to the specification of an implentable system, in particular because the users' needs are quite underspecified or beyond the reach of current technology; conversely, designers can also suggest potentially useful functionalities that users might have overlooked. In such cases, iterative and incremental development offers a more flexible approach, by performing several cycles of inception-elaboration-construction-transition between the initial planning and the final deployment. Still, browser development did not proceed through a sequence of prototypes adding more and more functionalities, but specifications and prototypes have emerged gradually from a series of back-and-forth exchanges between users and developers.

In our view, it is the helix-shaped model shown in Figure 6 that best represents the iterations that have shaped the field of meeting browsing. The helix rotates through four sectors that form the horizontal plane in Figure 6), while making progress towards more specific products on the vertical axis. Although this axis is correlated with time, it is not necessary that all components of a browser progress at the same pace, therefore the axis is better characterized in terms of specificity and product completeness. The four sectors for each iteration correspond to the most important stages in the development of software prototypes, especially of research ones. These stages match the ISO 9126 recommendations [ISO/IEC, 2001] and the IBM Rational Unified Process [Kroll and Kruchten, 2003], and include requirements, analysis and design, implementation, and testing.[8] Based on the meeting browser lifecycle, the four sectors can be divided by two axes: the "people" axis goes from users to developers, while the "systems" axis goes from construction to evaluation. Hence, the four sectors of the helix are: requirements elicitation, design/implementation, performance evaluation (i.e. intrinsic evaluation of components by developers), and task-based evaluation (i.e. extrinsic evaluation of a product by users). The evaluation of complex systems such as meeting browsers, which often rely on AI-inspired technology to extract features from meeting recordings, is not so much a matter of testing/verification, but rather of quantifying the performance level with respect to an error rate which is inevitably non-zero for most human-communication-analysis processes.

As in previous iterative models, evaluation results from one iteration, which can also be viewed as more or less specific elicited requirements, are used to derive new specifications for design in the next iteration or loop. Each loop of the helix ends with a certain form of evaluation, which can be a proper form of evaluation, or at least some kind of analysis of the product of the loop, based on users' experience with a more or less fully implemented

[8] They exclude 'business modeling' and 'deployment' from the RUP list since we are not dealing with end-user products.

prototype. Depending on each iteration or loop, the evaluation methods that are used may vary considerably to match the technologies under evaluation: e.g., from browser prototypes over hand-crafted annotations, possibly even including a human Wizard, to systems that integrate standalone multimedia processing tools.

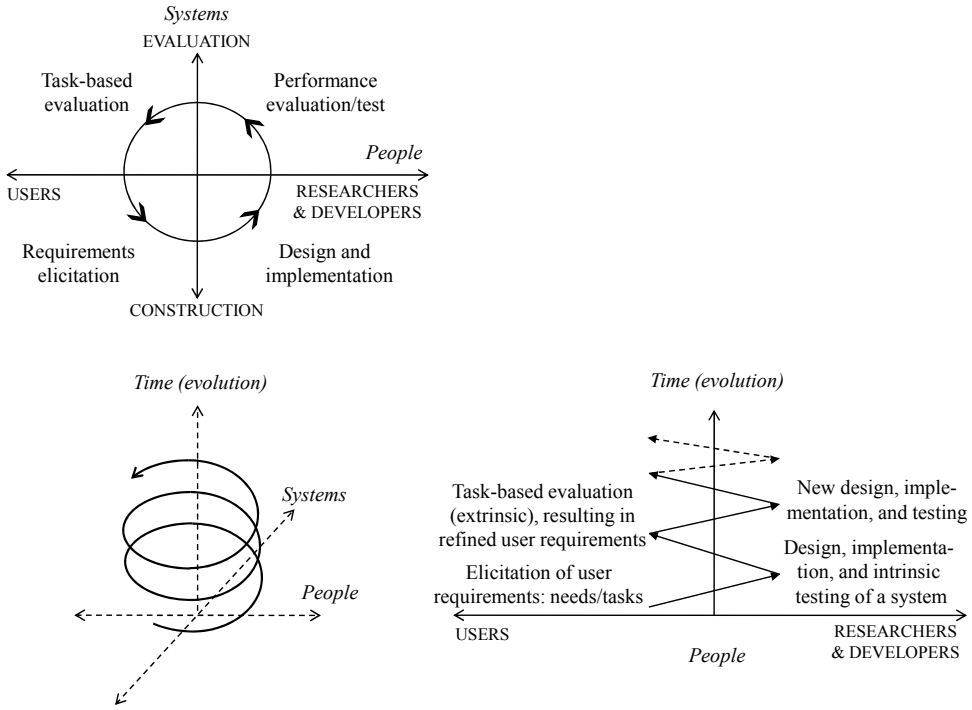

Figure 6: Software development process for meeting browsers: the helix model.

The helix model shares a common vision with Boehm's spiral model [Boehm, 1986], as they both represent the need for repeated iteration through the same key phases. However, in the spiral model, the emphasis is on risk analysis for the completion of a large project, and the loops of the spiral are incremental and intend to mitigate risks with respect to the final product, whereas in our case the goal is to develop and refine an open-ended set of functionalities. In particular, the sectors of Boehm's spiral represent different activities than ours (elaborate objectives, identify and resolve major sources of risk, define and elaborate the product, plan the next cycle), with the four helix stages all concentrated into one sector. However, the four

loops exemplified by Boehm [Boehm, 1986, Fig. 2] are not far from the four stages we will now overview, though presumably for larger projects more than four loops may be completed.

Research on meeting browsers carried out within our consortia have gone through four main iterations of the helix, as shown in Table 5. Work has cycled through these four levels sometimes simultaneously, and not always in strict sequential development. Each activity had components in each sector of the horizontal plane, and pushes forward our knowledge of meeting browsing tools one level up the vertical axis. From the least "implemented" browsers, i.e. requiring manual processing, to the most automated ones and further to the products, prototypes served different goals.

In a first phase, several studies were aimed at eliciting user requirements (using interviews and questionnaires to focus groups, i.e. requirements elicitation) and at the same time at studying the potential technology in order to design efficient and useful technologies for the tasks to be supported. The results included sets of meeting browsing tasks, databases of queries to meeting archives, and evaluation meant in fact the statistical analysis of user queries (to infer user requirements). Progressing towards more specified prototypes, a Wizard-of-Oz study was carried out, evaluated both through performance measures and behaviour analysis. In a third phase, functional research prototypes of meeting browsers and assistants have been implemented, partly based on the findings of the first phase, cas cited above. These more or less automated browser enabled quantitative user evaluations through the BET task-based procedure, and other efficiency and usability metrics for browsers, and as such to assess the usefulness of specific multimodal processing components. Finally, the most efficient and useful technology reached the level of end-user product, but for a slightly different task, as explained above: conference browsing. This enables to transfer the consortium know-how to a realistic application and were subject to field studies, for which evaluation is mainly done by estimating customer satisfaction for the products.

10 Conclusion

We have overviewed the main achievements related to the requirements and design of meeting browsers, carried out by two large consortia over eight years. The resulting picture of the software development process departs considerably from the waterfall model, according to which users have the primary role of formulating requirements for a task, and developers then attempt to design software satisfying them. Instead, specifications and prototypes emerged gradually from a series of exchanges between users and developers.

The analyses presented here show that, on the one hand, user requirements for meeting browsing cannot constitute a rigid, set-in-stone specifi-

Iteration	Methods	Outcomes	Assessment
1	Interviews, questionnaires	Meeting browsing tasks and databases of queries	Statistical analyses
2	Wizard-of-Oz experiments	Archivus	Behavior analysis and performance measures
3	Research prototypes of meeting browsers	Speedup, Overlap, JFerret, TQB, JFriDoc, FaericWorld, ViCoDe	Precision, speed, and other task-based efficiency and usability metrics
4	End-user products for presentation browsing	Klewel, SMAC	Commercial success, customer satisfaction

Table 5: Four iterations of the software process for meeting browsers.

cation, but depend greatly on how subjects are prompted to respond, and must be gradually focused towards a specifiable and implementable task. On the other hand, it has appeared that trusting only technology providers to evaluate the usefulness of their technology was unrealistic, leading to never-ending debates in which each provider tries to prove the utility of their own approach.

Our experience during the eight years of existence of our consortia, with literally hundreds of researchers collaborating together, has shown that user-driven and technology-driven approaches should be combined. Jumping back-and-forth from the users' to the developers' perspective has enabled us to gradually focus on the fact-finding task, providing both an applicative framework to develop innovative technologies and a reliable benchmark to evaluate their usefulness in a user-oriented setting.

Acknowledgments

The work presented in this paper has been supported by the Swiss National Science Foundation (SNSF) through the IM2 National Center of Competence in Research on Interactive Multimodal Information Management (http://www.im2.ch), and by the European Union through the AMI and AMIDA Integrated Projects (http://www.amiproject.org). The authors are grateful to their colleagues from the AMI and IM2 consortia for the long-standing and fruitful cooperation.

References

Gregory D. Abowd, Elizabeth D. Mynatt, and Tom Rodden. The human experience. *IEEE Pervasive Computing*, 1:48–57, 2002.

Marita Ailomaa, Miroslav Melichar, Martin Rajman, Agnes Lisowska, and Susan Armstrong. Archivus: a multimodal system for multimedia meeting browsing and retrieval. In *COLING/ACL 2006 Interactive Presentation Sessions*, pages 49–52, Sydney, 2006.

AMI. Meeting browser evaluation report. Deliverable D6.4, AMI Integrated Project FP6 506811 (Augmented Multi-party Interaction), December 2006.

Motoei Azuma. Square: The next generation of the ISO/IEC 9126 and 14598 international standards series on software product quality. In *ESCOM 2001 (12th European Software Control and Metrics Conference)*, pages 337–346, London, 2001.

Satanjeev Banerjee, Carolyn Rose, and Alexander I. Rudnicky. The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In *Proceedings of INTERACT 2005 (10th IFIP TC13 International Conference on Human-Computer Interaction)*, LNCS 3585, pages 643–656, Rome, 2005.

Enrico Bertini and Denis Lalanne. Total Recall survey. Technical report, University of Fribourg, Department of Computer Science, August 2007. URL http://diuf.unifr.ch/people/lalanned/Articles/TR-survey-report0807.pdf.

Nigel Bevan. International standards for HCI and usability. *International Journal of Human-Computer Studies*, 55:533–552, 2001.

Barry Boehm. A spiral model of software development and enhancement. *ACM SIGSOFT Software Engineering Notes*, 11(4):14–24, 1986.

Matt-M. Bouamrane and Saturnino Luz. Meeting browsing: State-of-the-art review. *Multimedia Systems*, 12:439–457, 2007.

Jean Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007.

Jean Carletta, Simone Ashby, Sébastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The AMI Meeting Corpus: A pre-announcement. In Steve Renals and Samy Bengio, editors, *Machine Learning for Multimodal Interaction II*, LNCS 3869, pages 28–39. Springer-Verlag, Berlin/Heidelberg, 2006.

Anita Cremers, Inge Kuijper, Peter Groenewegen, and Wilfried Post. The Project Browser: Supporting information access for a project team. In *Proceedings of HCII 2007 (12th International Conference on Human-Computer Interaction), Human Computer-Interaction, Part IV*, LNCS 4553, pages 571–580, Beijing, 2007.

Hoa Trang Dang, Diane Kelly, and Jimmy Lin. Overview of the trec 2007 question answering track. In *TREC 2007 (16th Text REtrieval Conference)*, NIST Special Publication 500-274, Gaithersburg, MD, 2007.

Bruno Dumas, Rolf Ingold, and Denis Lalanne. Benchmarking fusion engines of multimodal interactive systems. In *ICMI-MLMI 2009 (11th International Conference on Multimodal Interfaces and 6th Workshop on Machine Learning for Multimodal Interaction)*, pages 169–176, Cambridge, MA, 2009a.

Bruno Dumas, Denis Lalanne, and Rolf Ingold. HephaisTK: A toolkit for rapid prototyping of multimodal interfaces. In *ICMI-MLMI 2009 (11th International Conference on Multimodal Interfaces and 6th Workshop on Machine Learning for Multimodal Interaction)*, pages 231–232, Cambridge, MA, 2009b.

Laila Dybkjær, Niels Ole Bernsen, and Wolfgang Minker. Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication*, 43(1-2):33–54, 2004.

Florian Évequoz and Denis Lalanne. "I thought you would show me how to do it" – studying and supporting PIM strategy changes. In *ASIS&T 2009 Personal Information Management Workshop*, Vancouver, BC, 2009.

Philip N. Garner, John Dines, Thomas Hain, Asmaa El Hannani, Martin Karafiat, Danil Korchagin, Mike Lincoln, Vincent Wan, and Le Zhang. Real-time ASR from meetings. In *Interspeech 2009 (10th Annual Conference of the International Speech Communication Association)*, pages 2119–2122, 2009.

Dafydd Gibbon, Inge Mertins, and Roger K. Moore, editors. *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation.* Kluwer Academic Publishers, Dordrecht, 2000.

Julio Gonzalo, Paul Clough, and Alessandro Vallin. Overview of the CLEF 2005 Interactive Track. In Carol Peters and al., editors, *Accessing Multilingual Information Repositories (CLEF 2005 Revised Selected Papers)*, LNCS 4022, pages 251–262. 2006.

ISO/IEC. *ISO/IEC 9126-1:2001 (E) – Software Engineering – Product Quality – Part 1:Quality Model.* International Organization for Standardization / International Electrotechnical Commission, Geneva, 2001.

Alejandro Jaimes, Kengo Omura, Takeshi Nagamine, and Kazutaka Hirata. Memory cues for meeting video retrieval. In *Proceedings of CARPE 2004 (1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences)*, pages 74–85, New York, NY, 2004.

Per Kroll and Philippe Kruchten. *The Rational Unified Process Made Easy: A Practitioner's Guide to the RUP.* Addison-Wesley Professional, Harlow, 2003.

Denis Lalanne and Rolf Ingold. Documents statiques et multimodalité : l'alignement temporel pour structurer des archives multimédias de réunions. *Document numérique*, 8(4):65–89, 2004.

Denis Lalanne and Stéphane Sire. Analysis of end-user requirements: sample queries. Technical report IM2.AP, IM2 NCCR (Interactive Multimodal Information Management), 2003 2003.

Denis Lalanne, Dalila Mekhaldi, and Rolf Ingold. Talking about documents: revealing a missing link to multimedia meeting archives. In *Document Recognition and Retrieval XI, IS&T/SPIE's Annual Symposium on Electronic Imaging*, pages 82–91, San Jose, CA, 2004.

Denis Lalanne, Agnes Lisowska, Eric Bruno, Mike Flynn, Maria Georgescul, Maël Guillemot, Bruno Janvier, Stéphane Marchand-Maillet, Mirek Melichar, Nicolas Moenne-Loccoz, Andrei Popescu-Belis, Martin Rajman, Maurizio Rigamonti, Didier von Rotz, and Pierre Wellner. The IM2 multimodal meeting browser family. Technical report, IM2 Swiss National Center of Competence in Research (Interactive Multimodal Information Management), March 2005 2005.

Denis Lalanne, Laurence Nigay, Philippe Palanque, Peter Robinson, Jean Vanderdonckt, and Jean-François Ladry. Fusion engines for multimodal interfaces: a survey. In *ICMI-MLMI 2009 (11th International Conference on Multimodal Interfaces and 6th Workshop on Machine Learning for Multimodal Interaction)*, pages 153–160, Cambridge, MA, 2009.

Quoc Anh Le and Andrei Popescu-Belis. Automatic vs. human question answering over multimedia meeting recordings. In *Proceedings of Interspeech 2009 (10th Annual Conference of the International Speech Communication Association)*, pages 624–627, Brighton, UK, 2009.

Agnes Lisowska. Multimodal interface design for the multimodal meeting domain: Preliminary indications from a query analysis study. Technical report IM2.MDM-11, IM2 NCCR (Interactive Multimodal Information Management), November 2003 2003.

Agnes Lisowska, Andrei Popescu-Belis, and Susan Armstrong. User query analysis for the specification and evaluation of a dialogue processing and retrieval system. In *Proceedings of LREC 2004 (4th International Conference on Language Resources and Evaluation)*, pages 993–996, Lisbon, 2004.

Agnes Lisowska, Mireille Bétrancourt, Susan Armstrong, and Martin Rajman. Minimizing modality bias when exploring input preference for multimodal systems in new domains: the Archivus case study. In *Proceedings of CHI 2007 (ACM SIGHI Conference on Human Factors in Computing Systems)*, pages 1805–1810, San José, CA, 2007.

Stéphane Marchand-Maillet and Eric Bruno. Collection guiding: A new framework for handling large multimedia collections. In *AVIVDiLib 2005 (First Workshop on Audio-visual Content And Information Visualization In Digital Libraries)*, Cortona, Italy, 2005.

Miroslav Melichar. *Design of Multimodal Dialogue-based Systems*. PhD thesis, 4081, EPF Lausanne, School of Computer and Communication Sciences, 2008. URL http://library.epfl.ch/theses/?nr=4081.

Thomas P. Moran, Leysia Palen, Steve Harrison, Patrick Chiu, Don Kimber, Scott Minneman, William van Melle, and Polle Zellweger. "i'll get that off the audio": A case study of salvaging multimedia meeting records. In *Proceedings of CHI 1997 (ACM SIGCHI Conference on Human Factors in Computing Systems)*, pages 202–209, Atlanta, GA, 1997.

Gabriel Murray, Steve Renals, and Jean Carletta. Extractive summarization of meeting recordings. In *Interspeech 2005 (9th European Conference on Speech Communication and Technology)*, pages 593–596, Lisbon, Portugal, 2005.

Gabriel Murray, Thomas Kleinbauer, Peter Poller, Steve Renals, Jonathan Kilgour, and Tilman Becker. Extrinsic summarization evaluation: A decision audit task. In Andrei Popescu-Belis and Rainer Stiefelhagen, editors, *Machine Learning for Multimodal Interaction V (Proceedings of MLMI 2008, Utrecht, 8-10 September 2008)*, LNCS 5237, pages 349–360. Springer-Verlag, Berlin/Heidelberg, 2008.

Gabriel Murray, Thomas Kleinbauer, Peter Poller, Tilman Becker, Steve Renals, and Jonathan Kilgour. Extrinsic summarization evaluation: A decision audit task. *ACM Transactions on Speech and Language Processing*, 6(2):1–29, 2009.

Anselmo Penas, Pamela Forner, Richard Sutcliffe, Alvaro Rodrigo, Corina Forascu, Inaki Alegria, Danilo Giampiccolo, Nicolas Moreau, and Petya Osenova. Overview of

respubliqa 2009: Question answering evaluation over european legislation. In *CLEF 2009 (Cross-Language Evalution Forum Workshop)*, Corfu, Greece, 2009.

Andrei Popescu-Belis and Maria Georgescul. TQB: Accessing multimodal data using a transcript-based query and browsing interface. In *Proceedings of LREC 2006 (5th International Conference on Language Resources and Evaluation)*, pages 1560–1565, Genova, 2006.

Andrei Popescu-Belis, Philippe Baudrion, Mike Flynn, and Pierre Wellner. Towards an objective test for meeting browsers: the BET4TQB pilot experiment. In *Proceedings of MLMI 2007 (4th Workshop on Machine Learning for Multimodal Interaction)*, LNCS 4892, pages 108–119, Brno, 2008a.

Andrei Popescu-Belis, Erik Boertjes, Jonathan Kilgour, Peter Poller, Sandro Castronovo, Theresa Wilson, Alejandro Jaimes, and Jean Carletta. The AMIDA Automatic Content Linking Device: Just-in-time document retrieval in meetings. In Andrei Popescu-Belis and Rainer Stiefelhagen, editors, *Machine Learning for Multimodal Interaction V (Proceedings of MLMI 2008, Utrecht, 8-10 September 2008)*, LNCS 5237, pages 273–284. Springer-Verlag, Berlin/Heidelberg, 2008b.

Andrei Popescu-Belis, Jonathan Kilgour, Peter Poller, Alexandre Nanchen, Erik Boertjes, and Joost de Wit. Automatic content linking: Speech-based just-in-time retrieval for multimedia archives. In *SIGIR 2010 (33rd Annual International ACM SIGIR Conference on Research and Development on Information Retrieval), Demonstration Session*, Geneva, 2010.

Wilfried Post and Mike Lincoln. Developing and evaluating a meeting assistant test bed. In Andrei Popescu-Belis and Rainer Stiefelhagen, editors, *Machine Learning for Multimodal Interaction V (Proceedings of MLMI 2008, Utrecht, 8-10 September 2008)*, LNCS 5237, pages 338–348. Springer-Verlag, Berlin/Heidelberg, 2008.

Wilfried Post, Erwin Elling, Anita Cremers, and Wessel Kraaij. Experimental comparison of multimodal meeting browsers. In *Proceedings of HCII 2007 (12th International Conference on Human-Computer Interaction), Human Interface, Part II*, LNCS 4558, pages 118–127, Beijing, 2007.

Wilfried Post, Mirjam Huis In't Veld, and Sylvia van den Boogaard. Evaluating meeting support tools. *Personal and Ubiquitous Computing*, 12:223–235, 2008.

Maurizio Rigamonti, Denis Lalanne, Florian Évéquoz, and Rolf Ingold. Browsing multimedia archives through intra- and multimodal cross-documents links. In *Proceedings of MLMI 2005 (2nd Workshop on Machine Learning for Multimodal Interaction)*, LNCS 3869, pages 114–125. Edinburgh, UK, 2006.

Maurizio Rigamonti, Denis Lalanne, and Rolf Ingold. FaericWorld: Browsing multimedia events through static documents and links. In *Proceedings of Interact 2007 (11th IFIP TC13 International Conference on Human-Computer Interaction), Part I*, LNCS 4662, pages 102–115, Rio de Janeiro, 2007.

Ian Sommerville. *Software engineering*. Addison Wesley, Harlow, 9th edition, 2007.

Gokhan Tür et al. The CALO Meeting Assistant system. *IEEE Transactions on Audio, Speech and Language Processing*, 18(6):1601–1611, 2010.

Nikos Tsourakis, Agnes Lisowska, Pierrette Bouillon, and Manny Rayner. From desktop to mobile: Adapting a successful voice interaction platform for use in mobile devices. In *SiMPE 2008 (3rd ACM MobileHCI Workshop on Speech in Mobile and Pervasive Environments)*, Amsterdam, 2008.

Simon Tucker and Steve Whittaker. Accessing multimodal meeting data: Systems, problems and possibilities. In Samy Bengio and Hervé Bourlard, editors, *Machine Learning for Multimodal Interaction*, LNCS 3361, pages 1–11. Springer-Verlag, Berlin/Heidelberg, 2005.

Ellen M. Voorhees. The TREC question answering track. *Natural Language Engineering*, 7(4):361–378, 2001.

Ellen M. Voorhees. Overview of the TREC 2003 question answering track. In *TREC 2003 (12th Text REtrieval Conference)*, NIST Special Publication 500-255, pages 54–68, Gaithersburg, MD, 2003.

Ellen M. Voorhees and D. Tice. The TREC-8 question answering track evaluation. In *TREC-8 (8th Text REtrieval Conference)*, NIST Special Publication 500-246, pages 83–106, Gaithersburg, MD, 1999.

Alex Waibel, Michael Bett, Michael Finke, and Rainer Stiefelhagen. Meeting browser: Tracking and summarizing meetings. In Denise E. M. Penrose, editor, *Proceedings of the Broadcast News Transcription and Understanding Workshop*, pages 281–286, Lansdowne, VA, 1998.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *ACL/EACL 1997 (35th Annual Meeting of the Association for Computational Linguistics)*, pages 271–280, Madrid, 1997.

Pierre Wellner, Mike Flynn, and Maël Guillemot. Browsing recordings of multi-party interactions in ambient intelligent environments. In *CHI 2004 Workshop on "Lost in Ambient Intelligence"*, Vienna, Austria, 2004.

Pierre Wellner, Mike Flynn, and Maël Guillemot. Browsing recorded meetings with Ferret. In Samy Bengio and Hervé Bourlard, editors, *Machine Learning for Multimodal Interaction*, LNCS 3361, pages 12–21. Springer-Verlag, Berlin/Heidelberg, 2005a.

Pierre Wellner, Mike Flynn, Simon Tucker, and Steve Whittaker. A meeting browser evaluation test. In *Proceedings of CHI 2005 (ACM SIGCHI Conference on Human Factors in Computing Systems)*, pages 2021–2024, Portland, OR, 2005b.

Steve Whittaker, Patrick Hyland, and Myrtle Wiley. Filochat: Handwritten notes provide access to recorded conversations. In *Proceedings of CHI 1994 (ACM SIGCHI Conference on Human Factors in Computing Systems)*, pages 271–277, Boston, MA, 1994.

Steve Whittaker, Simon Tucker, Kumutha Swampillai, and Rachel Laban. Design and evaluation of systems to support interaction capture and retrieval. *Personal and Ubiquitous Computing*, 12(3):197–221, 2008.

Zhiwen Yu and Yuichi Nakamura. Smart meeting systems: A survey of state-of-the-art and open issues. *ACM Computing Surveys*, 42(2):8:1–16, 2010.

Klaus Zechner. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485, 2002.