AMIDA

Augmented Multi-party Interaction with Distance Access

`http://www.amidaproject.org/`

Integrated Project IST–033812

Funded under the 6th FWP (Sixth Framework Programme)

Action Line: IST-2005-2.5.7 Multimodal interfaces

Deliverable D6.3: HCI Evaluation of Prototype Applications

Due date: 30/09/2008 **Submission date:** 31/10/2008
Project start date: 1/10/2006 **Duration:** 36 months
Lead Contractor: Idiap Research Institute **Revision:** 1

Project co-funded by the European Commission in the 6th Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

D6.3: HCI Evaluation of Prototype Applications

Abstract:

This deliverable reports on the evaluations that we have performed on AMIDA end-user technologies during the second year of the project. This includes early-stage evaluations of the automatic content linking (ACLD) and user engagement and floor control (UE-FCD) "major" demos, in the form of three group discussions and one questionnaire-based study. Two discussions, run formally as focus groups with potential users, were based on screenshots of our demonstrations and on mock-ups. One group drew participants from the military sector, and the other was more general. A less formal group discussion was held at one of our Community of Interest events, and was based on very early basic demonstrations to make clear the general concepts behind our vision. In addition to this material evaluating the major demos, the deliverable also includes component technology evaluations for systems showcasing decision detection and summarization.

Contents

1 Introduction

AMIDA development is structured as follows. There are a range of underlying technologies, such as automatic speech recognition, dialogue act classification, and focus of attention tracking that as a set underpin any possible use of meeting archives or analysis of an ongoing meeting. Discussions with the user community, in the shape of the companies that make up the Community of Interest, and requirements capture from the project's first year have brought the project to the choice of two major demonstrations that show our vision for what this technology can provide. These two demos are the automatic content linking device (ACLD) and the user engagement and floor control demo (UEFCD), and they are described in detail in AMIDA D6.7. These demonstrations are subject to the complete software development cycle, from requirements capture, through development, to system evaluation. At this stage in the project, they are still in development, with system evaluation scheduled for the project's final year. In addition to these major demonstrations, where there is a promising application showcasing a particular component technology, we perform a system evaluation to show that the component technology brings benefits over what was available before AMIDA.

This deliverable reports on the evaluations that we have performed on AMIDA end-user technologies during the second year of the project. This includes two sets of early-stage evaluations of the ACLD and the UEFCD "major" demos, in the form of discussions with potential users. Two discussions, run formally as focus groups, were based on screenshots of our demonstrations and on mock-ups describing a surrounding technological context and advanced, unimplemented functionalities. The other was a less formal group discussion held at one of our Community of Interest events, and was based on very early basic demonstrations of the ACLD and the "Mobile Meeting Assistant" (MMA), which is a particular form for supporting user engagement and floor control designed for deployment on mobile phones. As well as these group discussions, we include one more formal questionnaire-based evaluation of the MMA. In addition to this material evaluating the major demos, the deliverable also includes component technology evaluations for systems showcasing summarization and decision detection.

2 First "major demo" focus group

The first focus group looking at the major ACLD and UEFCD demonstration concepts involved three facilitators and four participants who work in the military sector. We ran a focus group for this sector because it is of business importance to one of the project partners and because it presents specific constraints that could affect utility of AMIDA technologies. In the military setting, most meetings are face-to-face in order to ensure security, although experts are sometimes consulted at a distance. There are often few technological means available to the meeting participants, since meetings can be held in, for instance, temporary buildings, tents, or even vehicles. Because military organizations are rigidly hierarchal, communication lines are long, which can cause extra difficulties for fast and efficient communication.

2.1 Method

The focus group took two hours, divided as follows:

10 min. Preliminaries: basic background; explanation of procedure.

10 min. Discussion to share experiences of planning and holding both face-to-face and remote meetings.

20 min. Introducton to the two demos.

15 min. Discussion of needs, desires, and functionality.

15 min. Break.

35 min. Discussion of visualization, interaction, use.

5 min. Evaluation of the focus group.

The demos were introduced first by describing a setting in which team communication would be difficult, but new AMIDA technology might help. In the scenario, a geographically dispersed team has to communicate and exchange information using ICT that provides a virtual project environment. The environment provides support when someone is on the road (remote participation, limited bandwidth, distraction, limited visual attention, privacy), when someone enters the meeting late (catching up functionality), when someone is participating remotely (support on engagement, floor control, turn taking, back channeling, focus of attention), but also when the meeting is in need of information (content linking functionality) or in need of an expert (content linking, availability of the expert to the meeting, catching up). Then the ACLD and UEFCD concepts were described as possible partial solutions for providing such a virtual project environment, by talking around screenshots from early versions. For the ACLD, this was augmented with a graphical mock-up of potential future functionality that added more information about people in the organization and their expertise and participation in past meetings.

2.2 Results: Meetings in the military content

In the military context, most meetings are run to a rigid schedule with fixed speaking times. This makes it easy to predict when an agenda item is due, which makes it possible for participants to enter the meeting room at a later time. Daily briefings are required as a kind of one-way communication to pass essential information up the military hierarchy. People who are not present expect to be given the information from the briefing. In staff and section meetings, there will often be "back-bench" participants that the speakers consult when they are unsure on specific points.

In addition to these highly planned meetings, some military meetings are the more interactive and involving "think tank" or brainstorming activities. These bring together people with a shared interest, each of whom has specific expertise and their own information sources.

The group saw both advantages and disadvantages arising from having teams communicate when they were geographically dispersed. The advantages were not having to deal with the logistics of being in an operational area, with fewer people present in dangerous areas, as well as being able to draw in more experts including having experts available to more than one team at once. The disadvantages were that it is easy to understand a situation when one is there "on the ground", that distance communications were less stable and less secure, and that if people were at a distance it would be harder to be sure where they were and when they would be available for discussion.

2.3 Results: UEFCD

The group saw many possible useful functionalities that might aid user engagement and floor control. They favoured some kind of support for taking the floor, such as a push-to-talk button for the remote participant that would alert whoever was chairing the meeting from the main room. They also favoured giving remote participants feedback about whether the others are listening to them and when they are being addressed, the latter perhaps by turning on a small light. In this sector, lateness is seen as a fault, however, and so the main use of the technology that permits remote participation would be for inviting in experts. For those who were meant to be at the meeting but have failed to arrive, it would be used not for participating remotely, but for listening in to "catch up" on what is being missed. For this reason, it would be useful to be able to make notes as preparation for when one arrives. Listening in is also useful for back-bench participants, who would want to give feedback immediately on the subject under discussion.

In terms of the interface itself, the group favoured some visible means for identifying the current speaker, but otherwise, they feel that audio and display of powerpoint slides is sufficient information. They did consider showing the focus of attention by enlarging the representation for the speaker being attended, and perhaps some kind of movement to focus on the person being talked about. They did not see the need for emotion visualization or other advanced types of information. They also consider possible visualizations for when voices are raised. In the current interface, they found the visualization of the meeting participants distracting, with the focus of attention arrow moving to often and with its meaning unclear; it looked like an indication of the current speaker. They thought this could be corrected by only indicating who is being attended when the attention of the group lingers.

2.4 Results: ACLD

The group had some concerns about potential unintended side effects of the technology. They thought it might encourage a "I can afford to be late" mentality - although in this context, meetings are so tightly scheduled that people usually know exactly when they need to be there - and that unless it was possible to block sensitive material from being accessible for content linking, people might feel they were under too close, "Big Brother"-style observation.

The group thought that it would be useful to look back at the documents or Powerpoint presentations used during briefings, and to allow the back-bench that wasn't present to see

them as well, but because of the nature of these meetings, not at fragments of the briefings themselves. All documents should be identified by author, version number, an indication of whether this is the most recent version, the date, and the time. The group thought that in order to avoid distracting the user, the content linking needed to work autonomously in the background. However, this does not preclude using personal keywords.

The content linking concept was seen as adding value mostly for the individual meeting participant, by helping them retrieve information that has been forgotten and catch up with parts of a meeting missed through inattention or non-attendance. Access to personally held information was seen as useful, but so was sharing personal linked content with the group, as long as it would be possible to separate private and common spaces.

In terms of advanced functionality, the group wanted to be able not only to retrieve decisions, but also the assumptions and arguments that have lead to the decision. This gives more background information on the decision and can be of help when a decision has to be worked out by another team then the planning team or higher management. The system should be usable for after-action reviewing. A chat function for keeping in contact with the back-bench would be useful, as would being able to search for available experts on a given topic. New information relevant to a user's personal keywords or expertise could cause them to be sent a notification. The group also discussed an "off-topic" detection that would warn the chair when a discussion was drifting, in order to improve efficiency. In general, modern ICT and working methods were seen as appealing to the new generation of employees.

3 Second "major demo" focus group

The second focus group looking at the major ACLD and UEFCD demonstration concepts involved three facilitators and four participants with general experience of working in teams within a large organization.

3.1 Method

The focus group took one hour, divided as follows:

5 min. Preliminaries: basic background; explanation of procedure.

15 min. Introduction to the idea of using technology to support dispersed groups and discussion of the problems of remote meetings.

20 min. The Automatic Content Linking Device: introduction and discussion.

20 min. The User Engagement and Floor Control Device: introduction and discussion.

The introductory material used for this focus group was similar to that for the military focus group, but in a more compressed timeframe, and with a specific emphasis on the problem for remote groups of "cold starts", or the difficulty of engaging in informal "chit-chat", especially before a meeting, as would ordinarily happen when face-to-face.

3.2 Results: "cold start"

The focus group recognized the phenomenon of the "cold start", commenting that meetings held using ICT are more formal than they would be if held face-to-face because of the lack of "rich" communication. Informal communication does happen during remote meetings, but it takes more time. The technology particularly hinders jokes.

3.3 Results: UEFCD

The group discussed how people who wish to speak take the floor during a face-to-face meeting. There is a heavy reliance on nonverbal cues and "barging in" by just starting to talk. Use of technology makes this difficult. When two people start speaking at once in a remote meeting, then they often both stop, wait, and start talking together again. Remote meetings are thought especially irritating to extraverts, since they end up interrupting people unintentionally. Because the cues that someone else wishes to speak are often missed, people speak for longer, especially in audio-only meetings. The status cues that help participants determine when to speak, or how much, are much less apparent if they are not face-to-face.

It was important to group members that participants know what they look like in the interface when the other participants look at them. They wished participants to be shown in real size, and thought that the current visualization was too artificial and distracting - they preferred either some simpler display or a small but real video of the meeting room. Like the other focus group, they misread the green focus of attention arrow as a speaker indicator, and also commented that they were unclear why the camera switches viewpoints.

Also like the other group, they thought it useful to be able to see a Powerpoint presentation remotely and to have help knowing who was speaking, but for the latter, they suggested that 3D audio might help. Unlike the other group, saw the utility for a range of other kinds of information besides just speaker information. This included simple position in the room, role, and expertise. It also included visualizations for participant state. For instance, they thought it would be useful to know which participants were "zoned-out", and also which participants were having trouble understanding the discussion, since often people are reluctant to interrupt unless they know others share their problem.

3.4 Results: ACLD

The focus group thought the general concept of content linking useful for groups that have a large information-base from which to draw material, but suggested that content be linked only on demand, not continually. They thought it crucial that the system have access to all of the information that a participant might be able to find themselves. They wished for differentiation in how different information sources are presented, with familiar visualizations. For instance, Powerpoint presentations could look like slides, books like their covers with titles, and so on. They indicated that it would be crucial to be able to add one's own keywords to the system, and were concerned that keywords alone could not be used to find the right information. Other search aids might be the people involved

and meeting metadata such as who was at a meeting and where it was held.

They thought that adding an emotional layer to the information retrieval could help improve its value. For meetings, it is useful for the system to give some insight into who has the floor most. It is also useful in a series of meetings to understand what topics re-occur, when a discussion is being repeated, and also which topics never get discussed because they are forgotten. Tag clouds for topics can help to support this understanding. This kind of functionality would take the concept from information presentation towards intervention in the current group discussion.

The group was split on the question of who should have control of the content linking device; some members thought only the chair should, or some kind of process supervisor, and others thought that all participants should be given equal access.

The group commented that information from one's own peer group feels reliable, and that there is scope for using the content linking concept to support the team in a more community-oriented way. For instance, it might be possible to "whisper" individual discovery of a relevant resource, or some other information, to a colleague. Information about what content others have found relevant is valuable, and could be used to change what an individual sees. That is, rather than seeing a content linking device as a tool for individuals, it can be seen as a community tool where the results depend on what people do or share online.

4 Community of Interest discussion of the Automatic Content Linking Device

The first version of the AMIDA Automatic Content Linking Device was submitted to users and feedback was gathered and analyzed systematically. This section presents the methods and results of this analysis. The description of the system itself is available in Deliverable D6.7 AMIDA (2008) and in a paper presented at MLMI 2008 Popescu-Belis et al. (2008). To summarize, the ACLD relates ongoing discussions to potentially relevant documents or other pieces of information, and can be used in two scenarios.

1. *'Just-in-time' retrieval:* participants to a meeting are constantly made suggestions about documents (including excerpts of previous meetings) that are potentially relevant to the ongoing discussion. Participants are free to ignore them, or to start using them to enhance the discussion.

2. *Document/speech alignment for meeting browsers:* users of a meeting archive can view the recordings of previous meetings augmented with related documents, regardless of whether the participants to the meeting referred to them explicitly or not.

The execution tests of the ACLD have been satisfactory: the communication between the modules using the Hub works smoothly, and the logs show proper connection, sending and receival of annotation triples. The documents that are retrieved contain the expected words and keywords (this is easier to check on a static HTML representation of the result rather than the real-time one) and the functionalities planned for the first version are available.

Feedback from potential users and customers was obtained during several presentations of the ACLD, and most importantly as comments received at the AMIDA Community of Interest Workshop (Martigny, February 4–5, 2008) and at the AMIDA Know-how and Knowledge Transfer Day (Utrecht, September 10, 2008). The ACLD was part of an interactive program divided into "breakout sessions", in which it received the visit of about thirty representatives of companies from the AMIDA COI[1], which are active in the field of meeting technology. The sessions lasted 30' each, starting with a presentation of the ACLD and followed by questions and feedback from the audience (notes were taken). In addition, the ACLD was shown at a regional technology transfer event during the evening of February 4, 2008, providing additional feedback.

The received feedback can be analyzed into three main categories, plus a fourth "bucket class".

1. Graphical layout of the interface
    - the frame where the document names are displayed is too small with respect to the rest – a lot of space is lost in the interface – use a larger part of the screen for detecting documents;
    - display a larger overview of each document (e.g. as in Internet Explorer's "Quick Tabs" overview (Ctrl+Q)) in the main window, without keeping history data (i.e. past 30-second intervals);
    - avoid using too many mouse clicks: the content of documents should be more visible from the start;
    - colour code the nature of documents, and/or their relation to meetings;
    - make dates more visible.
2. Document repository
    - include documents that are known to users (from previous meetings) but also documents that are unknown to them: the value of the interface would then be to inform users about documents that they might not know of, not only the ones from previous meetings;
    - include shared documents but also non-shared ones for individual users (e.g. my emails) – increase personalization, for individual use: each user could have their own document repository and their own list of keywords;
    - include websites among documents, e.g. using a Google search (for instance, news sites, blogs, or websites of competitors) – without increasing too much the size of the repository;
    - allow for various document categories: project-specific, company-specific, external, websites, etc.
3. Additional functionalities

[1] See http://www.amiproject.org/business-portal/about-ami/project-partners/community-of-interest/vendors.

- after a meeting, email to all participants a list of (pointers to) the documents that were ”touched” (consulted) during the meeting
- allow group use but also individual use during a meeting. For instance, allow each participant to use their own version, and record the interest of each of them in specific documents, then show to everyone information about the group use of the document repository (which ones are most clicked?);
- implement relevance feedback: refine search technique based on user behavior (i.e. return more documents similar to those that are frequently consulted);
- detect specifically the similarities with previous discussions and alert uses that they already had this discussion before;
- represent the keywords as a cloud with emphasis varying with their frequency (see `http://www.quintura.com` for an example).

4. Varia

   - give the system more knowledge about the context of a meeting, and retrieve only documents related to that context (use word sense disambiguation over the query words) – improve accuracy of search;
   - both online and offline scenarios look interesting (online: meeting assistant; offline: meeting browser);
   - this application could lead to a cultural shift in the way meetings are organized: meetings could be divided into two short sessions, with an interval in the middle during which participants could consult the documents that were retrieved by the CLD;
   - frequent questions: who else is doing that?, has anyone else tried this?
   - infrequent question: how do you evaluate the quality of the result?

The feedback obtained during this user-centric evaluation is being used to produce the second version of the ACLD, which will be ready by end 2009.

The *performance evaluation* of the ACLD application is the topic of future work. One can of course test the performance of the retrieval system in terms of precision and recall, but this requires the definition of a ground truth document set for each time interval of a meeting, which is the main difficulty for such an evaluation. Several approaches of the ACLD performance evaluation are described in Deliverable D6.7.

5 Community of Interest Discussion of the Mobile Meeting Assistant

The *Mobile Meeting Assistant (MMA)* is a graphical interface for accessing remote meetings in real time from mobile devices, described in full detail in Deliverable D6.7 AMIDA (2008) and in a demo paper presented at MobileHCI 2008 Matena et al. (2008). For this version we focus on accessing annotated meetings of AMI corpus (Carletta et al., 2006).

Two main user interfaces (2D and 3D view) are used to show the current communication situation in the meeting room. The real time aspect is based on streaming recorded meetings, because of lack of algorithms and of hardware that would be able to produce necessary annotations in real time.

Feedback from COI workshop participants mainly concerns ideas for future improvements or developments. This feedback was obtained during the "breakout sessions", in which the MMM received the visit of about twenty representatives of companies from the AMIDA COI[2], which are active in the field of meeting technology. The sessions lasted 30' each, starting with a presentation of the ACLD and followed by questions and feedback from the audience. In addition, the system was demonstrated at a regional technology transfer event later during the first day, providing additional feedback.

1. Add more interfaces among which the user can choose, some of them possibly very abstract (position of participant corresponding to company hierarchy etc.). Let users choose their own avatars, possibly also used in other programs (e.g. `http://www.wee-mee.com`). Let users define their own meeting rooms (simple form / using image / select from a list).
2. Add access to information about other persons involved in the meeting.
3. Allow document sharing.
4. Integrate actual photos/camera streams.
5. Include visualization of longer-term meeting features (social network).
6. Keep the interface very simple.
7. The display of slides in real-time on the mobile phone could be turned very easily into a product.
8. The system could also be used for purely remote meetings, to create a feeling of meeting room.
9. Test it on real device.
10. Make a desktop version with more features.

The MMA appears thus quite intuitive to use, and seems especially useful to someone who is new to the meeting participants; however, 2D graphical conventions could be made clearer, and the 3D representation seems unnecessarily complex to some users. In both cases, slides should be made more visible, as well as the identity of the speakers. Ideally, more subtle body language, such as signs of disagreement and agreement, should be conveyed.

More realism was also required, e.g. by including actual photos of users, or at least letting them choose their own avatars. Accessing information about the other participants as well as meeting documents was another suggestion, while the slide capture itself appeared to be a good candidate for a commercial product. Of course, a realistic system running on a physical phone is a primary objective, bringing the audio stream to the user via the data stream (using VoIP), and synchronizing it with annotations and with the graphical representation of the meeting.

[2]See `http://www.amiproject.org/business-portal/about-ami/project-partners/community-of-interest/vendors`.

An important development will be the converse representation of the remote participant into the meeting room, because people in the meeting also need to improve their understanding of his/her presence beyond pure speech. At this point, the system could also be extended to support purely remote meetings, so that it creates the feeling of a meeting room using virtual reality.

6 Evaluating the MMA through a Pilot User Study

A small-scale user study was performed with 13 subjects, all of whom use information technologies every day and have a university degree in computer science. The subjects were given a demo of the MMA application running on an emulator in real-time, with a video recording of the meeting playing on second computer, for a duration of 5 minutes (meeting IS1008a from the AMI Corpus). They had then the possibility to interact with the application, for instance to change the interface or the viewing angle, for a maximum duration of 5 minutes.

The subjects answered a questionnaire shortly after the demonstration and had a possibility to add personal comments. Some questions required them to rate various aspects of the MMA device, as well as possibilities for future implementations, and other questions dealt with their own needs for a remote meeting assistant. Numeric ratings are coded from 1 to 5, 1 being best and 5 worst.

6.1 User Study: Questions and Raw Answers

We summarize below the questions that were asked in the user study, and the answers obtained from 13 subjects in [square brackets].

Q.1 Experience: please give us your opinion on the application that you just have tested. Rate from 1 (you like a lot) to 5 (you don't like at all) the following interfaces: 2D [AVG=2.3]; 3D standard [AVG=2.2]; 3D funny [AVG=2.3]; 3D advanced [AVG=2.7].

Q.2 Interface and colors used in the main demo were: too simple [8%] / just fine [84%] / too complicated [8%].

Q.3 Please rate following and planned features from 1 (you like a lot) to 5 (you don't like at all): who is speaking when/to whom [AVG=1.7]; entered room/left meeting alert [AVG=2.2]; possibility to create/upload personalized avatars [AVG=2.4]; head orientation [AVG=2.5]; focus of attention [AVG=2.1]; full-screen slide preview [AVG=2.0]; you are expected to speak! alert [AVG=2.0]; accurate representation of where people are sitting [AVG=2.8]; textual transcript of the meeting [AVG=2.5].

Q.4 Think about the possible use of this application in your perspective. What sort of meeting would you like to participate remotely in: business: yes [77%] / no [0%] / don't know [23%]; design/technical meeting: yes [85%] / no [15%] / don't know [0%]; personal (family, friends): yes [38%] / no [46%] / don't know [16%].

Q.5 Where would you use it? In the office: yes [69%] / no [23%] / don't know [8%]; On the train/airplane: yes [69%] / no [23%] / don't know [8%]; At the airport/train station:

yes [84%] / no [16%] / don't know [0%]; In a car: yes [33%] / no [59%] / don't know [8%]; Other situations: yes [33%] / no [0%] / don't know [67%].

Q.6 What limitations do you see for use on the road? Select as many as you wish: environment too disturbing to be on meeting [7/13]; privacy concern [4/13]; screen/device too small [7/13]; time available [1/13]; attention available (driving the car, switching trains) [10/13].

Q.7 Overall rating. Rate from 1 (you like a lot) to 5 (you don't like at all) the following questions. How do you like the idea? [AVG=1.5] How do you like our current approach? [AVG=1.9] Is it easy to understand who is speaking? [AVG=2.2] I could imagine being more engaged in the meeting. [AVG=2.5]

Q.8 When ready and adapted to your needs would you be using it? Regularly [2/13] / sometimes [9/13] / never [2/13].

Q.9 Desktop version. Would you prefer a desktop version? yes [61%] / no [23%] / don't know [16%]. Would you also use a desktop version? yes [92%] / no [8%] / don't know [0%].

Q.10 Additional comments. Give us a personal feedback, hint or request a feature.

6.2 Synthesis of the results

The subjects judged the MMA very positively, as they liked the concept (1.5/5) and the present approach (1.9/5) – if a positive answer is counted if 1 or 2 was the response, 92% liked the general idea and 85% liked the current approach. They would use such an application "sometimes" (9 out of 13), mainly for design/technical meetings (11 out of 13) or business meetings (10 out of 13), but less for personal meetings (5 out of 13). They would mainly use the application while waiting at the train station or at the airport (10 out of 13), in the office or on a train/airplane (9 out of 13 both). The main limitations for use in such conditions is the available attention if the user must do something else (e.g. go to a gate or catch a train), the small size of the screen, and noise from the environment (7 out of 13 both).

In terms of the users' experience, they seem equally satisfied with the 2D and the 3D interfaces (2.3/5 and 2.2/5)[3]. The interface and color schemes are at the appropriate level of complexity (11 out of 13). The most appreciated information is "who is speaking when/to whom" (1.7/5), followed by the full-screen slide preview (2.0/5), the focus of attention (2.1/5) and head orientation (2.5/5). Possible features to be added in the future have been rated similarly: "you are expected to speak" alert seems the most desired one (2.0/5), followed by "enter/leave room" (2.2/5), use of personalized avatars (2.4/5), and speech transcript (2.5/5)[4].

[3] We observed in fact a variety of acceptance for each interface when doing the study: while an important proportion of users prefer a representation which is close to reality – such as a 3D representation – while others prefer a simple 2D representation, or even funnier alternatives. Several users suggested the possibility to upload their own avatars for each participant.

[4] In other words, regarding the selected media channels and annotations, the user study showed that 77% of users would like to see who is speaking when and to whom, 69% of users were interested to know what participants are looking at (focus of attention). Another strongly requested feature was the full screen slide preview (by 69% of participants). Finally the expected to speak feature was appreciated by 73% of

Finally, most of the subjects would also use a desktop version of the MMA (11 out of 13), and some would even prefer it (8 out of 13), a fact that meets some of the explicit suggestions received from industrial partners.

6.3 User Study: Suggestions

From the user study questionnaire (Q10), we reformulate and synthesize from actual users' feedback the following ideas. They are grouped into two (inter-related) categories: evaluation of demonstrated system, and suggestion for the future.

Evaluation from user study:

1. The demo looks quite intuitive to use.
2. The demo is of great use for someone who is relatively new to the meeting participants, but when participants are familiar with each other, this graphical information looks superfluous, and speech could be enough.
3. 2D graphical conventions are not clear enough (ask a good graphic designer).
4. It is uncertain that 3D adds anything new, but visual complexity. (*Another user:* 3D interfaces look more like a gadget than a really useful feature.)
5. Subtle body language is not conveyed – reactions such as signs of disagreement and agreement.
6. The interfaces are too primitive to be convincing (realism is not required, but richness and clarity are).
7. The slides should be more readable (having an unreadable black board is more disturbing than helping).
8. The most useful informations (who is speaking and the slides) should remain the most visible compared to the other features.

Future developments suggested by the user study:

1. Give people in the meeting a means to see the presence of the remote user virtually (otherwise the remote user is only watcher/listener). To be active in the meeting, they need to be present virtually.
2. Record the meeting so that one can track back on interesting parts, and use speech transcription for indexing.
3. Try to spread this application widely, not only for business use but also for personal communication.
4. Solve also the challenges of a real phone (using data and voice at the same time).
5. Can speaker and VFOA annotations be done in real time? What if two persons speak simultaneously?

participants of the poll.

7 Evaluation of summarization technologies using a decision audit task

7.1 Decision audit task

The decision audit task involved a user reviewing previously held design team meetings in order to determine how a given decision was reached. This task could be accomplished by determining the final decision, the alternatives that were previously proposed, and the arguments for and against the various proposals. This task was chosen because it represented a key application, that of aiding corporate memory, the storage and management of a organization's knowledge, transactions, decisions, and plans. An organization may find itself in the position of needing to review or explain how it came to a particular position or why it took a certain course of action. We hypothesize that this task will be made much more efficient if multimodal meeting recordings—and the means to browse the recordings—are available, along with their summaries.

The decision audit represents a complex information need that cannot be satisfied with a simple one-sentence answer. Relevant information will be spread across several meetings and may appear at multiple points in a single discussion thread. Because the decision audit does not only involve knowing *what* decision was made but also determining *why* the decision was made, the person conducting the audit will need to understand the evolution of the meeting participants' thinking and the range of factors that led to the ultimate decision. For a particular decision audit task, the decision itself may be a given. Because the person conducting the decision audit does not know which meetings are relevant to the given topic, there is an inherent relevance assessment task built into this overall task. Their time is limited and they cannot hope to scan the meetings in their entirety and so must focus on which meetings and meeting sections seem most promising. In contrast to the task-based evaluation, previously reported, this evaluation was an individual rather than group-based task.

Each participant in the decision audit task was first given a pre-task questionnaire, relating to background, computer experience and experience in attending meetings, followed by the task instructions. The portion of the instructions detailing the specific task read as follows:

> *We are interested in the group's decision-making ability, and therefore ask you to evaluate and summarize a particular aspect of their discussion.*
>
> *The group discussed the issue of separating the commonly-used functions of the remote control from the rarely-used functions of the remote control. What was their final decision on this design issue? Please write a short summary (1-2 paragraphs) describing the final decision, any alternatives the participants considered, the reasoning for and against any alternatives (including why each was ultimately rejected), and in which meetings the relevant discussions took place.*

This particular information need was chosen because the relevant discussion manifested itself throughout the 4 meetings, and the group went through several possibilities before

designing an eventual solution to this portion of the design problem. The task itself involved reviewing a complete meeting series (ES2008) from the AMI corpus, comprising four related, sequential meetings.

After completing the decision audit task, participants answered a post-task questionnaire.

7.2 Summarisation

We evaluated five approaches summarisation using the decision audit approach:

- Baseline (**KAM**): choose the top 20 keywords (selected using the su.idf term weighting scheme Murray and Renals (2007))
- Automatic extractive summary of manual transcripts (**EAM**)
- Automatic extractive summary of automatic speech recognition (ASR) transcripts (**EAA**)
- Human-authored abstractive summaries of manual transcripts (**AMM**)
- Semi-automatic abstractive summaries of manual transcripts (**ASM**)

The extractive summarization was performed using a support vector machine (SVM) with radial basis functions (RBF) kernel to classify each DA as extractive or non-extractive, trained on the AMI labelled training data (90 scenario meetings) using 17 features from five broad feature classes: prosodic, lexical, length, structural and speaker-related. This system is described in greater detail in deliverable D5.2, and in Murray and Renals (2007). The lengths of the abstractive summaries varied between 30–40% of the original meeting length. For the ASR case the overall word error rate was 38.9%.

The human-authored abstractive summaries varied in length. Each abstractive sentence was normally also linked to one or more transcript DAs, making the experimental condition a hybrid of abstractive and extractive. Because this was a decision audit task and the abstractive summary provided in this condition had a "decisions" subsection, we considered this to be a high-quality gold-standard condition. The semi-automatic abstractive summaries Kleinbauer et al. (2007) used hand-annotated topic segmentation and topic labels, and detected the most commonly mentioned content items in each topic. A sentence was generated for each meeting topic indicating roughly what was discussed, and these sentences were hyperlinked to the actual DAs in the discussion. These summaries relied on manual transcripts, and so Condition EAA was the only ASR condition in this experiment. The Condition ASM summaries were only semi-automatic, since they relied on manual annotation of propositional content.

A meeting browser was constructed for each condition, built so as to exhibit as similar browser behaviour as possible across the conditions. Figure 1 shows an example of the browser interface for Condition AMM.

To evaluate the decision audit task, we analyzed three types of features: the answers to the users' post-questionnaires, human ratings of the users' written answers, and features extracted from the logs of mouse and keyboard activity relating to browsing behaviour

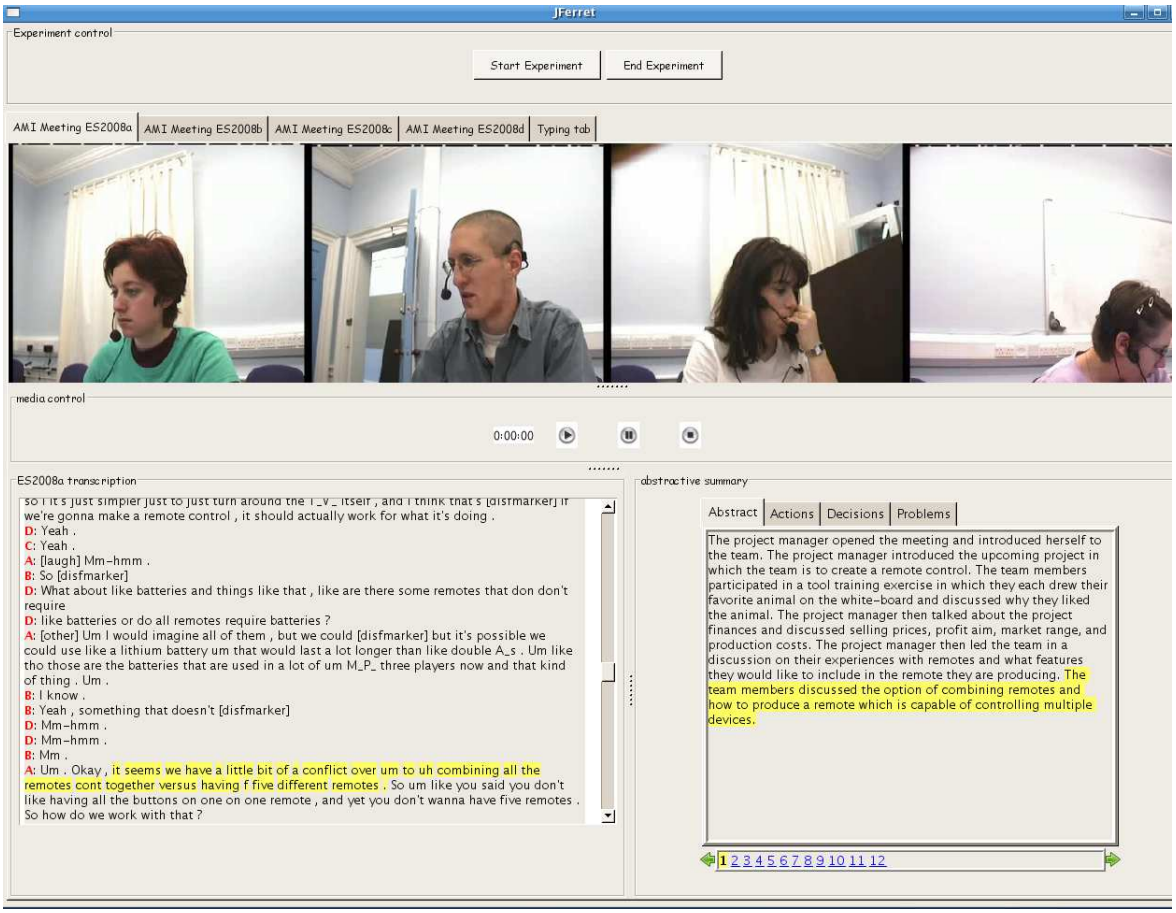

Figure 1: Condition AMM Browser

in the different conditions. Upon completion of the decision audit task, we presented each participant with a post-task questionnaire consisting of 10 statements with which the participant could state their level of agreement or disagreement via a 5-point Likert scale, such as *I was able to efficiently find the relevant information*, and two open-ended questions about the specific type of information available in the given condition and what further information they would have liked.

In order to gauge the quality of a participant's answer, we enlisted two human judges to do both subjective and objective evaluations. For the subjective portion, the judges first read through all 50 answers to get a view of the variety of answers. They then rated each answer using an 8-point Likert-scale on criteria roughly relating to the precision, recall and f-score of the answer. For the objective evaluation, three judges constructed a gold-standard list of items that should have been contained in an ideal summary of the decision audit. For each participant answer, they checked off how many of the gold-standard items were contained. The remainder of the features for evaluation were automatically derived from the logfiles. These features have to do with browsing and writing behaviour as well as the duration of the task.

Question	KAM	EAM	EAA	AMM	ASM
Q1: *I found the meeting browser intuitive and easy to use*	3.8	4.0	3.02^{AMM}	$4.3_{EAA,ASM}$	3.7^{AMM}
Q2: *I was able to find all of the information I needed*	2.9^{AMM}	3.8	2.9^{AMM}	$4.1_{KAM,EAA,ASM}$	3.0^{AMM}
Q3: *I was able to efficiently find the relevant information*	2.8^{AMM}	3.4_{ASM}	2.5^{AMM}	$4.0_{KAM,EAA,ASM}$	$2.65^{EAM,AMM}$
Q4: *I feel that I completed the task in its entirety*	2.3^{AMM}	3.1	2.3	3.2_{KAM}	2.9
Q5: *I understood the overall content of the meeting discussion*	3.8	**4.5**	3.9	4.1	3.9
Q6: *The task required a great deal of effort*	3.0	2.6_{EAA}	3.9^{EAM}	3.1	3.2
Q7: *I had to work under pressure*	3.3	**2.6**	3.3	2.7	3.1
Q8: *I had the tools necessary to complete the task efficiently*	3.1^{EAM}	$4.3_{KAM,EAA,ASM}$	3.0^{EAM}	4.1	3.5^{EAM}
Q9: *I would have liked additional information about the meetings*	3.0^{EAM}	2.0_{KAM}	2.4	2.6	2.7
Q10: *It was difficult to understand the content of the meetings using this browser*	2.1	$1.5_{EAA,ASM}$	2.7^{EAM}	2.0	2.3^{EAM}

Table 1: Post-Questionnaire Results
For each score in the table, that score is significantly worse than the score for any conditions in superscript, and significantly better than the score for any condition in subscript.

7.2.1 Results and discussion

Post-questionnaire results Table 1 gives the post-questionnaire results for each condition. For each score in the table, that score is significantly worse than the score for any conditions in superscript, and significantly better than the score for any condition in subscript. The only significant results listed are those that are significant at the level ($p<0.05$) according to non-paired t-test. Results that are not significant but are nonetheless unexpected or interesting are listed in boldface.

The gold-standard condition AMM scored best on many of the criteria, showing that human abstracts are an efficient way to survey and index into the content of a meeting. For example, participants in this condition found that the meeting browser was easy to use (Q1) and that they could efficiently find the relevant information (Q3).

A striking result is that not only were the manual extracts in condition EAM also rated highly on many post-questionnaire criteria, this condition was in fact the best overall for several of the questions. For example, on being able to understand the overall content of the meeting discussion (Q5) and having the tools necessary to complete the task efficiently (Q8), Condition EAM scored best.

However, it's clear that extracts of ASR output posed challenges that significantly decreased user satisfaction levels according to several of the criteria. For example, participants in Condition EAA found the browser less intuitive and easy to use (Q1), found it more difficult to understand the meeting discucssion (Q5) and used considerable effort to complete the task (Q6). On several criteria this condition rated the same or worse than the baseline Condition KAM, which uses manual transcripts.

Condition ASM incorporating semi-automatic abstracts generally rated very well in comparison with the gold-standard condition, scoring not significantly worse than Condition AMM on criteria relating to the ability to understand the meeting discussion and complete

the task (Q4 and Q5) and the effort required to complete the task (Q6 and Q7).

In general, the participants found the task to be challenging, as evidenced by the average answers on questions 4, 6 and 7. The task as designed required efficient navigation of the information in the meetings in order to finish the task completely and on time.

The gold-standard human abstracts were rated highly on average by participants in that condition. As mentioned earlier, this gold-standard condition was expected to do particularly well considering that it was a decision audit task and the abstractive summaries contain subsections that were specifically focused on decision-making in the meetings. The semi-automatic summaries (ASM) rated well in terms ease of use and intuitiveness, but slightly less well in terms of using the browser to locate the important information. It did consistently rate better than Conditions KAM and EAA, however.

The results of the post-questionnaire data are encouraging for the extractive paradigm in that the users seemed very satisfied with the extractive summaries relative to the other conditions. However, it is quite clear that the errors within an ASR transcript presented a considerable problem for users trying to quickly retrieve information from the meetings. While it has repeatedly been shown that ASR errors do not cause problems for these summarization algorithms according to intrinsic measures, these errors made user comprehension more difficult. For the questions relating to the effort required, the tools available, and the difficulty in understanding the meetings, Condition EAA was easily the worst, scoring even lower than the baseline condition. It should be noted however, that a baseline such as Condition KAM was not a true baseline in that it was working off of *manual* transcripts and would be expected to be worse when applied to ASR.

The ASR used in these experiments had a WER of about 39%; it is to be expected that these findings regarding the difficulty of human processing of ASR transcripts will change and improve as the state-of-the-art in speech recognition improves. The finding also indicates that the use of confidence scores in summarization is desirable. While summarization systems naturally tend to extract units with lower WER, the summaries can likely be further improved for human consumption by compression via the filtering of low-confidence words.

Human Evaluation Results - Subjective and Objective Table 2 gives the results for the human subjective and objective evaluations, formatted analogously to table 1. As before, Condition AMM scored best in the subjective as well as in the objective evaluation for most criteria. But we also observe that neither Condition ASM nor Condition EAM were significantly worse. However, the introduction of ASR had a measurable and significant impact on the subjective evaluation of quality. At the same time, what these findings together help illustrate is that automatic summaries can be very effective for conducting a decision audit by helping the user to generate a concise, complete high-quality answer.

For the objective human evaluation, the gold-standard condition scored substantially higher than the other conditions in hitting the important points of the decision process being audited. This indicates that there is much room for improvement in terms of automatic summarization techniques. However, Conditions EAM, EAA and ASM averaged much higher than the baseline Condition KAM. There is considerable utility in such automatically-generated documents. It can also be noted that Condition EAM was the best of the condi-

Criterion	KAM	EAM	EAA	AMM	ASM
Q1: *overall quality*	3.0^{AMM}	4.15	3.05^{AMM}	$4.65_{KAM,EAA}$	4.3
Q2: *conciseness*	$2.85^{EAM,AMM,ASM}$	4.25_{KAM}	3.05^{AMM}	$4.85_{KAM,EAA}$	4.45_{KAM}
Q3: *completeness*	2.55^{AMM}	3.6	2.6^{AMM}	$4.45_{KAM,EAA}$	3.9
Q4: *task comprehension*	$3.25^{EAM,AMM}$	$\mathbf{5.2}_{KAM,EAA}$	$3.65^{EAM,AMM}$	$5.25_{KAM,EAA}$	4.7
Q5: *participant effort*	4.4	$\mathbf{5.2}_{EAA}$	$3.7^{EAM,AMM,ASM}$	5.3_{EAA}	4.9_{EAA}
Q6: *writing style*	4.75	5.65_{EAA}	$4.1^{EAM,AMM,ASM}$	5.7_{EAA}	5.8_{EAA}
Q7: *objective rating*	4.25^{AMM}	7.2	5.05^{AMM}	$9.45_{KAM,EAA}$	7.4

Table 2: Human Evaluation Results - Subjective and Objective

For each score in the table, that score is significantly worse than the score for any conditions in superscript, and significantly better than the score for any condition in subscript.

tions with fully-automatic content selection (Condition ASM is not fully automatic).

Perhaps the most intriguing result of the objective evaluation is that Condition EAA, which uses ASR transcripts, did not deteriorate relative to Condition EAM as much as might have been expected considering the post-questionnaire results. What this seems to demonstrate is that ASR errors were annoying for the user but that the users were able to look past the errors and still find the relevant information efficiently. Condition EAA scored much higher than the baseline Condition KAM that utilized manual transcripts, and this is a powerful indicator that summaries of errorful documents are still very valuable documents. This relates to the previous findings of the SCANMail browser evaluation Hirschberg et al. (2001); Whittaker et al. (2002), in which participants were able to cope with the noisy ASR data.

An interesting question is whether participants' self-ratings on task performance correlated with their actual objective performance according to the human judges. To answer this question, we calculated the correlation between the scores from post-questionnaire Q4 and the objective scores. The statement Q4 from the post-questionnaire is "I feel that I completed the task in its entirety." The result is that there was a moderate but significant positive correlation between participant self-ratings and objective scores (pearson=0.39, $p<0.005$).

Figure 2 shows the relationship between the objective ratings and participant self-ratings for all 50 participants. While the positive correlation is evident, an interesting trend is that while there were relatively few people who scored highly on the objective evaluation but scored low on the self-ratings, there were a fair number of participants who had a low objective score but rated themselves highly on the post-questionnaire. A challenge with this type of task is that the participant simply may not have had a realistic idea of how much relevant information was out there. After retrieving four or five relevant items, they may have felt that they had completed the task entirely. This result is similar to the finding by Whittaker et. al Whittaker et al. (pear), where participants often felt that they performed better than they really did.

Interaction Log Results Table 3 gives the results for the interactive log evaluation, formatted analagously to the previous tables. An unexpected result was that the task duration (Q1) did not vary significantly between conditions. Because the task was difficult to complete in 45 minutes, most participants took all or nearly all of the allotted time, regardless of condition. The impact of the gold-standard Condition AMM is clear: partic-

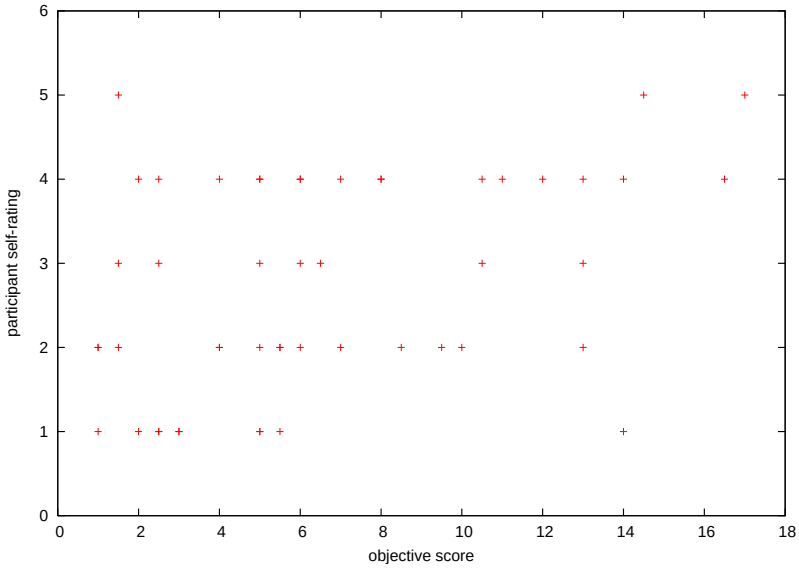

Figure 2: Objective Scores and Post-Questionnaire Scores

Feature	KAM	EAM	EAA	AMM	ASM
Q1: *duration*	45.4	43.1	45.4	45.42	43.2
Q2: *first typing*	16.25	13.9	17.14	8.61	10.22
Q3: *tabbing*	0.98	0.81_{AMM}	0.72_{AMM}	$1.4^{EAM,EAA}$	1.13
Q4: *perc. buttons clicked*	0.39	0.11	0.08	0.08	0.18
Q5: *clicks per minute*	1.33	**2.24**	1.47	1.99	0.83
Q6: *media clicks*	15.4_{EAA}	14.4_{EAA}	$40.4^{KAM,EAM,AMM}$	16.6_{EAA}	20.6
Q7: *click/writing corr.*	0.03	0.01	0.01	0.01	0.01
Q8: *unedited length*	1400	1602	1397	**2043**	1650
Q9: *edited length*	1251	1384	1161	**1760**	1430
Q10: *num. meetings*	3.9	4.0	3.9	4.0	4.0
Q11: *ave. writing timestamp*	0.68	0.73	$0.76^{AMM,ASM}$	0.65_{EAA}	0.65_{EAA}

Table 3: Interaction Log Results

For each score in the table, that score is significantly worse than the score for any conditions in superscript, and significantly better than the score for any condition in subscript.

ipants in this condition began writing their answer earlier (Q2), did not wait until the end to write the bulk of their answers (Q11), wrote longer answers (Q8) and had more time for editing their answers (Q9).

Perhaps the most striking finding from this analysis is the variation in how participants used the audio/video stream. In Conditions KAM, EAM, ASM and AMM, the number of media clicks (Q6) averaged around 14-20 per task. For Condition EAA, incorporating ASR, the average number of media clicks was 40.4, significantly higher than all other conditions with the exception of Condition ASM. Participants in Condition EAA relied much more on audio and video during this task. While they still used the summary dialogue acts to index into the meeting record (Q5), they presumably used the audio and video to disambiguate any ASR errors.

It is difficult to derive a single over-arching conclusion from the logfile results, but there were several interesting results on specific logfile features. Perhaps the most interesting was the dramatic difference that existed in terms of relying on the audio/video record when

using ASR. The average number of media clicks when using extractive summaries on manual transcripts was only just above 14, but when applied to ASR this number was over 40 clicks. This ties together several interesting results from the post-questionnaire data, the human evaluation data, and the logfile data. While the ASR errors seemed to annoy the participants and therefore affected their user satisfaction ratings, they were nonetheless able to employ the ASR-based summaries to locate the relevant information efficiently and thereby scored well according to the human objective evaluation. Once they had indexed into the meeting record, they then relied heavily on the audio/video record presumably to disambiguate the dialogue act context. It was not the case that participants in this condition used only the audio/video record and disregarded the summaries, as they clicked the content items more often than in Conditions KAM and ASM (Q5). Overall, the finding is thus that ASR errors were annoying but did not obscure the value of the extractive summaries.

7.2.2 Conclusion

Overall these results are very encouraging for the extractive summarization paradigm. Users find extractive summaries to be intuitive, easy-to-use and efficient, are able to employ such documents to locate the relevant information in a timely manner according to human evaluations, and users are able to adapt their browsing strategies to cope with ASR errors. While extractive summaries might be far from what people conceptualize as a meeting summary in terms of traditional meeting minutes, they are intuitive and useful documents in their own right.

Perhaps the most interesting result from the decision audit overall is regarding the effect of ASR on carrying out such a complex task. While participants using ASR find the browser to be less intuitive and efficient, they nonetheless feel that they understand the meeting discussions and do not desire additional information sources. In a subjective human evaluation, the quality of the answers in Condition EAA suffers according to most of the criteria, including writing style, but the participants are still able to find many of the relevant pieces of information according to the objective human evaluation. We find that users are able to adapt to errorful transcripts by using the summary dialogue acts as navigation and then relying much more on audio/video for disambiguating the conversation in the dialogue act context. Extractive summaries, even with errorful ASR, are useful tools for such a complex task, particularly when incorporated into a multi-media browser framework.

There is also the possibility of creating browsing interfaces that minimize the user's direct exposure to the ASR transcript. Since we have previously found that ASR does not pose a problem for our summarization algorithms, we could locate the most informative portions of the meeting and present the user with edited audio and video and limited or no textual accompaniment, to give one example.

8 Evaluation of decision detection technology

8.1 Introduction

Meetings are a critical aspect of most organizations. In meetings two or more people gather to discuss a topic, hoping to reach conclusions through the communication process. This process involves a shared goal and intensive oral arguments which provide rationales for individuals' points of view. Repositories of the audio-visual recordings of meeting dialogues constitute a valuable source of information for future training and group decision support Post et al. (2004); Romano and Nunamaker (2001). With the recent advances in recording and storage technologies, a rapidly growing number of meetings are being archived for later retrieval, and solutions are needed to help users better leverage the archived meeting recordings.

Standard meeting browsers, which come with typical information retrieval and playback facilities, help answer less than 20% of user queries Pallotta et al. (2007). This has led researchers to augment meeting browsers with additional plug-ins. For example, plug-ins that display topics and represent speaker roles and meeting states have been found to be effective for meeting information retrieval, helping users find the information they seek in 25% less time Banerjee et al. (2005).

In addition, recent organizational and user query studies have been conducted to identify what key aspects are missing from current meeting browsers Cremers et al. (2005); Lisowska et al. (2004); Pallotta et al. (2007). Pallotta et al. Pallotta et al. (2007) have highlighted the argumentation process and outcome as the most sought-after information, composing ***60%*** of common user queries. In particular, the argumentation outcomes, i.e., decisions, have been suggested as the most essential Pallotta et al. (2007); Post et al. (2004); Rienks et al. (2005); Romano and Nunamaker (2001); Wellner et al. (2005).

In this paper, we investigate whether using a plug-in that displays summaries of what the users want to know will help them find information from the archives more efficiently and effectively. Specifically, we evaluate the use of summary displays for the task of identifying the essential argumentation outcomes from previous meetings, a common practice for meeting preparation Pallotta et al. (2005); Rienks et al. (2005); Wellner et al. (2005). First, we experiment with a summary display that demonstrates what are indicative of overall discussion (which we will call "general-purpose summary display"). Murray Murray and Renals (2007) has evaluated this summary display and shown it rated well by the users who are asked to audit how a group came to a particular decision.

However, since the general-purpose summaries are often lengthy – Murray Murray and Renals (2007) extracted 10% of the meeting dialogue acts, they are expected to be less effective for the decision debriefing task than summaries that are focused on decisions. Therefore, in this paper, we also evaluate the use of a display that presents summaries that are tailored to user queries (which we will call "decision-focused summary display"). To the best of our knowledge, no user study has been performed to compare the effectiveness of using the general-purpose and the decision-focused summary display. Figure 3 exhibits how the two types of summaries are displayed alongside with the transcripts and audio-video recordings in a meeting browser.

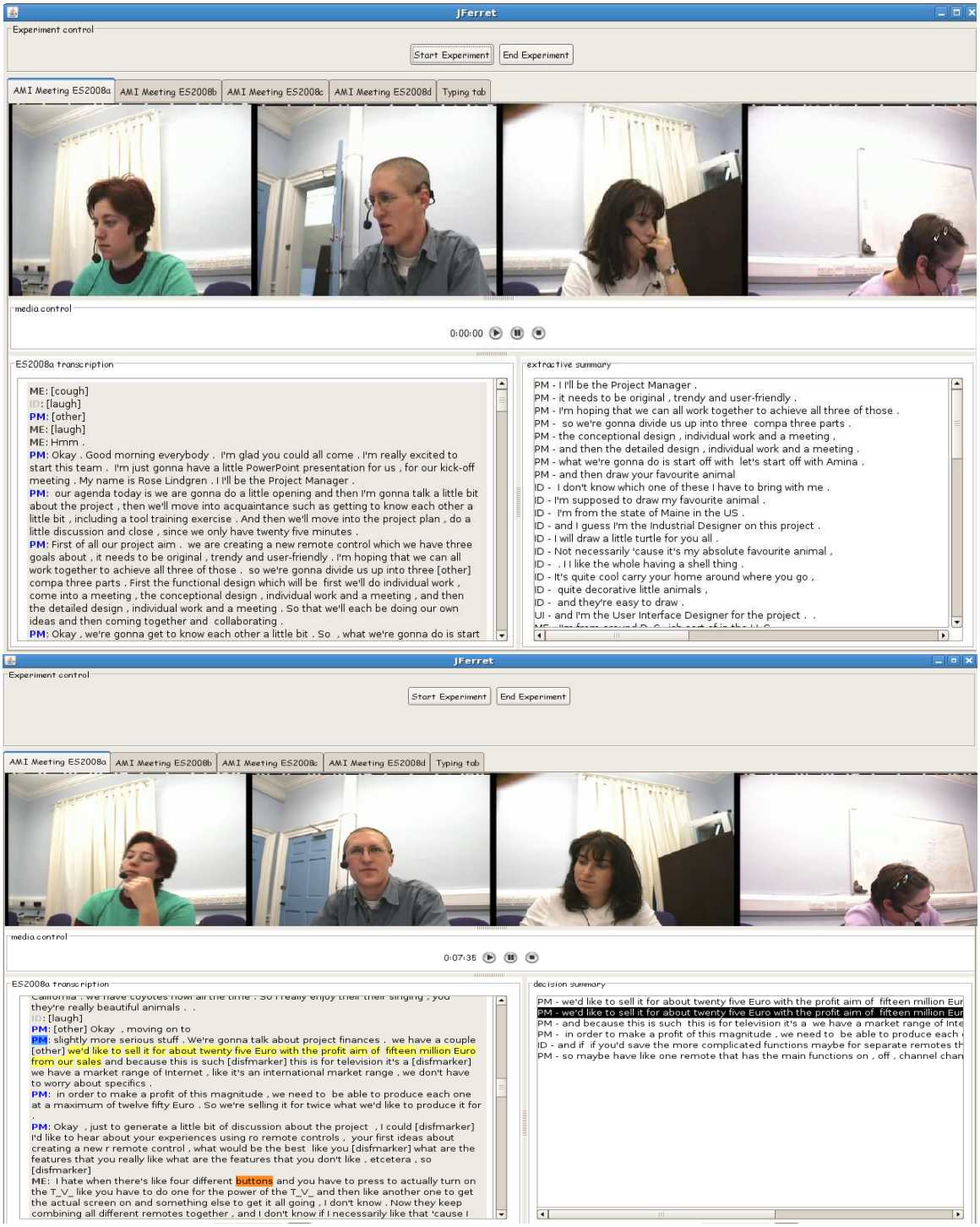

Figure 3: *Example AMI browsers. Both are composed of three plug-ins: the playback facility of audio-video recordings (top), the transcription display (lower left), and the extractive summary display (lower right). The two browsers differ only in the summary display plug-in, with the browser on the top demonstrating the general-purpose summary display and the one on the bottom demonstrating the decision-focused summary display.*

8.2 Related Work

To save time and human labor in generating meeting minutes, techniques have been developed to produce extractive summaries by distinguishing the informative dialogue units from the uninformative ones in meetings.

Traditionally, the extractive technique works well in text summarization, e.g., Mani and Bloedorn Mani and Bloedorn (1998) has found that users absorb information in summaries more quickly than in full text, despite some loss of accuracy. Text summarization commonly uses lexical information, such as counts of cue phrases, word co-occurrences, and tf*idf scores (or its variants), to rank the extract-worthiness of each unit Edmundson (1968); Kupiec et al. (1995); Teufel and Moens (2002). Some methods rely on orthographic cues (e.g., the position in text, title) and semantic information (e.g., the degree of connectedness in a semantic graph, co-references) Mani and Bloedorn (1998); Barzilay (2003). However, some of these features are not available for speech. Previous research in speech summarization remedies this problem by using other types of speech-specific information. For example, some researchers have combined lexical and prosodic information to perform summarization in speech genres such as broadcast news Koumpis and Renals (2001) and voice-mail Maskey and Hirschberg (2005).

Extractive techniques have also been applied to identify generically informative units that are reflective of overall meeting content Miekes et al. (2007); Murray et al. (2005); Zechner (2002). For example, Murray and Renals (2007) used prosodic information to identify the most informative meeting dialogue acts for general purpose extractive summaries.

Despite its effectiveness in the decision audit task, these often-lengthy general-purpose extractive summaries are insufficient for users who need a quick overview of all the information relevant to a particular query. In our own prior work Hsueh et al. (2007), we have followed previous work in query-driven summarization to select only those dialogue units that help fill in a decision-related template. The technique developed in this work identified lexical as well as multi-modal cues (e.g., gestures, head movements, prosody) that are predictive of decision-related dialogue acts.

8.3 Methodology

8.3.1 Task overview

As pointed out in many organizational studies, obtaining an overview of the decisions made in previous meetings is critical to the preparation of future meetings Pallotta et al. (2005); Rienks et al. (2005); Wellner et al. (2005). We hence use a "decision debriefing" task in this study to compare the two types of extractive summaries. The goal of this task is to summarize all the decisions made in a series of meetings.

We recruited 35 subjects (20 females and 15 males, ages from 18 to 44) during the period of two months in 2008 to perform this task. These subjects were recruited from the undergraduate and graduate program of distinctively diverse fields (e.g., history, medicine, chemistry, geography). They were asked to fill in a pre-questionnaire about their prior experience in computer use and meeting attendance. An experimenter then guided the subject through the procedure. In this evaluation, subjects were asked to go through one

AMI meeting series and to debrief the decisions for their upper management. The four meetings in the series are displayed in parallel so that the subjects could easily jump to the meeting recording they were interested in knowing more.

At the beginning of each session, the experimenter introduced the browser interface to the subject. The subjects were then free to browse through one pre-selected meeting recording (which is not used in the real experiment). They could take as much time they needed to familiarize themselves with the interface.

The main task is as follows:
"*In 45 minutes or less, write a report to summarize the decisions made in the four meetings for upper management*.[5]"

Because some subjects in a pilot study expressed the need to be reminded of the time remaining in the experiment, the experimenter signalled the subjects twice before the end of the experiment, once at 25 minutes and again at 40 minutes into the experiment. The subjects could also signal the experimenter to end the session if they finished the task early. During the session, all the user behaviors were recorded in the log files, and the user-generated decision minutes were logged separately.

At the end of each session, the experimenter asked the subjects to explain how they used the browser interface to find out about the decisions made in the meetings. Subjects were also asked to fill in a post-questionnaire about their perceived task success.

8.3.2 Meeting Corpus and Annotation

To obtain the annotations of decision-focused extractive summaries and gold standards for evaluating the user-generated minutes, we use the AMI meeting corpus Carletta (2006), in which meeting participants are required to make decisions as a group in a series of four product design meetings that are intended to imitate a typical product design cycle, starting with a kick-off meeting and ending with an evaluation meeting. To annotate all the meetings in this corpus, two groups of annotators were asked to go through a two-phase procedure:

- First, annotators were asked to navigate the recordings of one series of four meetings and to summarize the decisions made in these meetings into a list of "decision points" (as exemplified in Figure 4).
- Then, another group of three annotators were asked to go through the dialogue acts in each meeting one by one, and judge if they could be annotated as a "decision-related dialogue act (decision DA)", i.e., if they supported any of the decision points.

Decision point annotation

In the first phase of the annotation procedure, the set of decision points that were noted by *two ore more* annotators are used as the gold standard set of decision points. In the

[5]Our pilot study demonstrated that the decision debriefing task is straightforward enough to be completed in 45 minutes.

- The group decided not to define the target user group by a specific age range but simply by interest in fashion and simplicity.
- The remote will feature a locator function and large buttons.
- The remote will incorporate both simple and complicated functions, hiding the complicated functions from the main interface.
- The remote will be made to look fashionable.

Figure 4: *Example decision points of a product design meeting.*

meeting series used in this study, the meeting participants reached 6, 10, 8, and 6 decisions respectively.

Following is a synopsis of the original summaries that are in the form of bullet points (as in Figure 4).

- In the kick-off meeting (*A*), the entire group decided that the prototype design should be simple, keeping the everyday functions on one interface and more complicated functions on another;
- In the conceptual design (*B*) and detailed design (*C*) meetings, the group decided on the specific target group, the essential functions of the interface and the layout;
- In the wrap-up/evaluation meeting (*D*), the group decided on which prototype to choose and what functions to be eliminated from the prototype.

Decision-focused extract annotation

After each annotator finished their annotations, the ground truth "decision-focused extract" of each meeting (e.g., those used in the decision-focused summary display in Figure 3) is then generated by collecting the set of decision DAs that were extracted by *one or more* annotators.

An analysis of the decision-focused extract annotation shows that the annotators found on average four decision points per meeting and specified around two decision links for each decision sentence in the set of decision points. Overall, 554 out of 37,400 DAs (in a 50 meeting dataset) were annotated as decision DAs, accounting for 1.4% of all DAs in the data set and 12.7% of the original extractive summaries (which consist of the extracted DAs).

8.3.3 Meeting Browser Interface

The meeting browser (cf. Fig. 3) used in this evaluation is an enhanced version of the AMI meeting browser, which is designed to present additional annotations on top of the meeting recordings Carletta (2006). The enhanced version consists of three basic components: the audio-visual recording playback facility (top), the transcript display (lower left), and the extractive summary display (lower right).

Each subject is equipped with a headphone so that they can listen to the audio recordings whenever necessary. Users can play the audio-video recording from the beginning. Users

Independent Variable	**Levels**	**Dependent Variable**	**Factors**
Extract Type	General-purpose	Task effectiveness (User summary-based)	Ratio of correctly found decisions to all decisions in the model summary
	Decision-focused	Report quality (User summary-based)	Overall quality, completeness, conciseness, trustworthiness, style
		Perceived success (Post-questionnaire)	Ease of use, task completeness, decision coverage and comprehension (See Table 5)

Table 4: Experimental design.

who are interested in a particular decision DA can click on that "DA button" in the display and be led to the point where the DA was uttered in the dialogue. Each of the decision DA buttons in the summary display is time synchronized to the location of the decision DA in the audio-visual recording as well as in the transcript.

There are five tabs on the top of the browsing interface: (1) the first four tabs take users to each of the four meetings in the series chosen for display, and (2) the last tab is the "writing tab", where users are asked to type in their summaries. Users can switch between these tabs at will. During the experiment, a logging tool in the back-end records all the clicking and typing behaviors. With this log, we can analyze the use of the different components in the browser, such as the summary display and the audio-video playback facility, as well as the report typing behavior, e.g., how many characters were deleted, inserted, and substituted by each subject.

8.3.4 Experiment Design

Our research questions are concerned with the effect of automatic summary type on users' task performance and perceived success. Our hypothesis is that a more succinct, decision-focused summary (around 10% of the general purpose extractive summary) would help users obtain an overview of decisions more efficiently, prepare a meeting minute for upper management more effectively, and feel more confident in the meeting preparation work. In addition, we also test the impact of automation on the performance of the decision-focused summary. The subjects we recruited were randomly assigned into four groups. Each group was asked to accomplish the decision debriefing task, using one of the following summary displays embedded in the meeting browser:

- Baseline (AE-ASR): automatic general purpose extracts, automatic speech recognized (ASR) transcription[6].

[6]The ASR transcription used in this experiment was generated with a state-of-the-art program, which on

DV Factors	Post-questionnaire questions
Perceived ease of use (interface)	Q1: I found the meeting browser intuitive and easy to use.
Perceived ease of search	Q2: I was able to find all of the information I needed.
Perceived efficiency	Q3: I was able to efficiently find the relevant information.
Perceived task completeness	Q4: I feel that I completed the task in its entirety.
Perceived comprehension (general)	Q5: I understood the overall content of the meeting discussion.
Perceived task success (decision)	Q6: I was able to efficiently find the decisions.
Perceived task difficulty	Q7: The task required a great deal of effort.
Perceived pressure	Q8: I had to work under pressure.
Perceived system usefulness	Q9: I had the tools necessary to complete the task efficiently.
Perceived lack of support	Q10: I would have liked additional information about the meetings.

Table 5: Questionnaire-based measures of user perceived success and usability.

- AD-ASR: automatic decision-focused extracts[7], ASR transcription.
- AD-REF: automatic decision-focused extracts, manual transcription.
- Topline (MD-REF): manual decision-focused extracts, manual transcription.

We designed our experiment around the hypothesis that a decision-focused summary display benefits users more than a general-purpose display for accomplishing the decision debriefing task. The independent and dependent variables are shown in Table 4. The independent variables are tested between subjects.

The dependent variables are classified into three categories:

1. *Task effectiveness*: First of all, the user-generated decision minutes are evaluated against the gold standard decision points. Task effectiveness is measured by the percentage of the gold standard decision points that have been correctly listed in the user-generated decision minute ("decision hits").

2. *Report quality*: Different aspects of the user-generated decision minute quality are rated on 1-7 Likert scale. These aspects include the overall quality, completeness, conciseness, task comprehension, amount of effort spent in writing, trustworthiness, and writing style.

3. *User perceived success*: Finally, Table 5 lists the self-reported measures of the level of perceived success and usability, reported on 5-point Likert-scales in the post-questionnaire.

In addition to the three evaluation criteria, we have also developed a number of quantifiable measures to understand user behavior when given the different types of summary displays. These measures, which are computed from the log files, include task difficulty, task completeness, effort required, proportion of useful extracts, reading speed, productivity, usage of media, and usage of summary display. For example, the usage of summary

average recognizes words with 30%-40% error rate. Word error rate is a common metric of the performance of speech recognition systems. 30% to 40% error rate means that there is a 30%-40% chance for a word to be substituted, deleted, or incorrectly inserted.

[7]The automatic decision-focused extracts used in this experiment were generated by our state-of-the-art decision detection program Hsueh et al. (2007), which predicts decision DAs with 60%-70% agreement with the model summaries.

User Behavior	Measures
Task completeness	Number of meetings that have been read
First decision written	Time to type first character
Proportion of useful extracts	Number of content clicks, normalized by number of content buttons
Writing speed	Number of times switching to the writing tab, normalized by experiment length
Reading speed (Extract)	Number of content clicks, normalized by length of experiment
Productivity (by writing time-stamp)	Average time-stamp of insertions, normalized by experiment length
Productivity (by report length)	Number of words in user's report (edited)
Usage of media	Number of times user played the audio or video
Usage of extracts to correct writing	Number of content clicks in the preceding 2 minutes of a writing tab click

Table 6: Log file-based measures of task effectiveness.

display is measured by counting the number of decision DAs the user clicks on. The assumption is that the more DAs selected by the user, the more likely it is that the summary presents information that the user wants to know. A more detailed account of the log-based measures of user behaviors is presented in Table 6.

8.4 Results

In the context of the decision debriefing task, this study aims to answer the main question:

- Whether general-purpose extractive summaries–which extract generically important dialogue acts that reflect overall meeting content–could be improved by focusing on only the decision-related dialogue acts.

In addition, since we would like to know how much automation degrades the usefulness of the summary display, we also address the following two questions:

- Can the automatic decision-focused extracts help users achieve performance comparable to that obtained by navigating the manual extracts?
- Does operating on the transcription produced by automatic speech recognition (ASR) as opposed to manual transcriptions affect user performance significantly?

8.4.1 Effect of Summary Display Type on Decision Debriefing

In this section, we report the results of task effectiveness and report quality obtained from the analysis of the log files and the minutes. We also assess the user behavior in the use of the different types of summary display.

Task effectiveness analysis

The data analysis shows that the users on average yield more decision hits by using the decision-focused summary display than by using the general-purpose one (Figure 5). To determine whether the differences are statistically significant, an analysis of variance was performed. The meeting summary display type was found to have a significant main effect on task effectiveness ($F(3, 31) = 13.832; p < 0.001$). The best performing subject were able to use the Topline (manual decision extract, manual transcription) browser to find almost all the decision points.

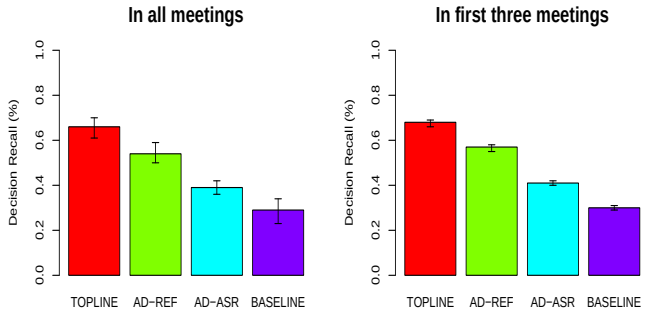

Figure 5: *Task effectiveness as the average ratio of the decisions that are correctly found by the subjects. These ratios are obtained from all meetings in the series (with a total number of 30 decision points) and from the first three meetings (with 24 decision points).*

Report quality analysis

A condition (4) x overall quality (5) analysis of variance on the decision minute ratings (Table 7) finds the meeting summary display type to also have a significant main effect on its overall quality ($F(3,31) = 3.324; p < 0.05$). With the additional information in the decision-focused summary display, the subjects are able to generate decision minutes of higher quality.

Criterion (1-7)	TOP	AD-REF	AD-ASR	BASE
Overall Quality	2.5	2.4	3.6	3.9
Completeness	3.1	2.9	3.8	3.4
Conciseness	2.4	2.7	2.6	3.4
Writing Style	2.6	2.1	3.3	3.4
Trustworthiness	1.9	2.0	1.8	2.4

Table 7: Quality assessment of the subjects' minutes. Results are obtained on a 7-point scale: the lower the score, the better the minute quality.

Perceived success analysis

The average ratings reported in the post-questionnaires (cf. Table 8) suggest that the decision-focused display is perceived to be easier to use ($F(3,31) = 4.819; p < 0.05$) and less demanding in the amount of effort required ($F(3,31) = 4.343; p < 0.05$). The subjects using the decision-focused display also find themselves able to retrieve the relevant information more efficiently ($F(3,31) = 8.710; p < 0.01$), and absorb the decisions made in the meetings more effectively ($F(3,31) = 4.714; p < 0.05$).

User behavior analysis

An ANOVA test on the log-based data reveals that, compared to the subjects in the baseline condition (automatic general-purpose extract, ASR transcription), subjects use a significantly higher proportion of the extracted decision DAs to write minutes ($F(3,31) = 9.878; p < 0.001$) and rely more on the extract contents to modify their minutes ($F(3,31) = 21.715; p < 0.001$). (See Table 9.)

Criterion (1-5)	TOPLINE	AD-REF	AD-ASR	BASELINE
Perceived ease of use (interface)	4.4	4.1	4.3	3.6
Perceived efficiency	3.9	3.4	3.6	3.3
Perceived comprehension (general)	4.6	4.6	4.1	4.1
Perceived task success (decision)	4.3	4.3	3.8	3.7
Perceived task difficulty	2.6	2.9	2.9	3.7
Perceived pressure	2.8	3.8	2.7	3.4
Perceived system usefulness	4.4	4.3	4.1	.4.1

Table 8: User perceived task success. Results are obtained on a 5-point scale (5 = agree strongly, and 1 = disagree strongly).

8.4.2 Decision-focused Extracts v.s. General-purpose Extracts

Given that the decision-focused summaries are more effective than the general-purpose summaries for the decision debriefing task, we wish to determine whether the effectiveness remains when a more error-prone automatically generated summary is used in the interface. To determine patterns that were not specified a priori, posteriori pairwise comparisons were performed.

First, we examined the decision hits across the conditions that use the automatic general-purpose display (BASELINE) and that used automatic decision-focused display (AD-ASR). The percentage of decision hits (as reported in Figure 5) shows that focusing on only the decision-related information results in greater task effectiveness–on average, increasing the number of decision hits over that yielded with the general-purpose display by 36%. Moreover, the decision minutes generated by the subjects who use decision-focused summaries tend to exhibit better overall quality and conciseness.

Further analysis of user behaviors reveal that the subjects still rely more on the decision-focused display to summarize meeting decisions, even when the summary contains ASR errors. The decision-focused display is found to significantly increase the use of the summary display ($p < 0.001$; Tukey's test), normalized frequency of switching to the writing tab ($p < 0.05$), and the usage of the summary display prior to writing correction ($p < 0.001$).

	TOPLINE	AD-REF	AD-ASR	BASELINE
Proportion of useful extracts	0.53	0.61	0.51	0.12
Usage of media	23.60	17.71	33.44	15.56
Usage of extracts to correct writing	6.84	1.54	0.93	0.66

Table 9: Task effectiveness measures based on user behavioral cues.

8.4.3 Automatically Generated Extracts v.s. Manual Extracts

The question that emerges naturally next is how much performance degradation results from replacing the manual summary with its automatic version (which contains 30%-40% inconsistencies with the ground truth). The answer would provide useful guidance for the design of meeting browsers, and may provide support for the development of automatic machinery for query-focused speech summarization.

To answer this question, we compared task effectiveness and report quality of the condition that uses manual decision-focused summaries (TOPLINE) to the one that uses automatically generated summaries (AD-REF). Although the overall quality of minutes in the two conditions does not differ significantly, the automatic extractive summaries have on average three fewer decision hits (21%).

To further understand whether the errors in the summary resulted in any systematic difference in user behavior, we examine the log files (cf. Table 9). We expected that users would prefer to use the meeting summary display to find decisions when the summaries are reflective of the actual decisions made. The post-hoc test results match the expectation: Using the automatic version of the summary (Column AD-REF) instead of the manual version (Column TOPLINE) significantly decreased the use of the summary display prior to writing correction (Tukey's test, $p < 0.01$).

However, this difference in task effectiveness and user behavior does not seem to affect the subjects' perceived success towards the task and ability to produce quality minutes: no significant difference was found in any of the subjects' ratings in the post-questionnaire and the minute quality ratings for the two conditions.

8.4.4 Effect of Transcription Type

Because our ultimate goal is to design a meeting browser that can be used as soon as a meeting ends, it is important to study whether operating the browser on error-prone automatic speech recognition (ASR) transcription (which contains 30%-40% errors) affects task effectiveness and report quality.

To examine the performance degradation caused by the ASR transcript display, post-hoc tests were also performed across the conditions that operate on ASR transcription (AD-ASR) and on manual transcription (AD-REF). The assessment results of report quality (cf. Bar AD-REF and AD-ASR in Fig. 5) suggest that displaying decision-focused summaries on manual transcripts helps the subjects find 39% more (on average, 4 to 5) decision hits than displaying the summaries based on ASR transcripts ($p < 0.01$).

Further analysis of the decision minute quality (cf. Table 7) shows that users who browse summaries on manual transcripts are likely to produce decision minutes of better overall quality and completeness. In addition, the more readable transcripts allow the subjects to allocate more of their time to absorbing relevant information, rather than understanding meeting content. In turn, the decision minutes generated by this group of users can be better appreciated by readers.

Examination of user behaviors (cf. Column AD-REF and AD-ASR in Table 9) also shows the transcription type to have effects on the usage of the summary display for writing

decision minutes ($p < 0.05$). The less helpful displays increase the level of perceived pressure ($p < 0.05$) reported by the subjects (cf. Row 6 in Table 8).

8.5 Discussion

Task Effectiveness	Report Quality	User Perception
Proportion Used ($F(3,31) = 9.878; p < 0.001$)	Overall quality	Easy to use ($F(3,31) = 4.819; p < 0.05$)
Use of extracts before writing	($F(3,31) = 3.324; p < 0.05$)	Effectively finding information ($F(3,31) = 8.710; p < 0.01$)
($F(3,31) = 21.715; p < 0.001$)		Required effort ($F(3,31) = 4.343; p < 0.05$)

Table 10: ANOVA results of task effectiveness for subjects across all four conditions.

The results of this study verify our experimental hypothesis. Displaying decision-focused summaries in the meeting browser helps users to obtain an overview of the decisions from multiple meeting recordings more effectively and efficiently than general-purpose summaries. The decision-focused summary, obtained by filtering out the dialogue acts irrelevant to decisions, was found to improve not only task effectiveness, but also the overall quality of the subjects' minutes. The users in the focused summary conditions read through a higher proportion of summary material to find relevant information and relied more on summaries to prepare and correct the decision minutes they wrote.

Having established the advantage of the decision-focused summary, our investigation further examined the impact of using automatically generated decision-focused summaries and operating a meeting browser on ASR transcripts. The first examination showed that, even when the displayed decision summaries were automatically generated, participants who used the decision-focused summary display still outperformed those who used the general-purpose summary display in the decision debriefing task. Although the automatic summary users did not achieve the same level of task effectiveness as those using manual summaries, they were able to produce decision minutes of similar quality.

One explanation for this could be that parts of the automatic summary correctly identified some of the decision points, and users leveraged the correct parts to find information relevant to these decision points for summarization. In fact, the user behavior analysis found task effectiveness (i.e., the number of decisions hits) and usage of the summary display (i.e., the proportion of the summary that was used to prepare and correct writing) to be significantly correlated (Spearsman's test; $r = 3.573$, $p < 0.001$).

From the post-experiment debriefings, we observed two main strategies adopted by users to find the decision points that were not clearly presented in the summaries: (1) Some users attempted to go through the extracted DAs in the summary display one by one looking for relevant information in the surrounding context in the transcript; (2) Others turned

to the audio-video recordings to find the missing decisions. The two coping strategies can be distinguished by their usage of media. Table 11 presents the proportion of subjects that have high and low usage of the audio video aids.

It appears that when the manual transcripts are in in the display, e.g., in the Topline and AD-REF conditions, the choice of strategy was based on individual differences, and a majority of the users preferred to use the decision-focused summaries rather than the audio-video aids. Yet when the error-prone ASR transcripts are in the display, e.g., in the AD-ASR and Baseline conditions, the choice of strategy was noticeably affected by the the type of summary display. Comparing Columns AD-ASR and Baseline in Table 11 illustrates that the AD-ASR users tended to make more usage of the audio-video recordings. This is because the ASR transcripts are difficult to understand by themselves, and it is therefore important to find additional hints from the the summary display; However, as the summaries presented in the AD-ASR display are often short and error-prone, the audio-video recordings are necessary for accomplishing the task.

Media Usage	**Topline**	**AD-REF**	**AD-ASR**	**Baseline**
Low (< 30)	70.0%	85.7%	44.4%	88.9%
High(>= 30)	30.0%	14.3%	55.6%	11.1%

Table 11: The proportion of subjects who had low and high usage of audio-video recordings: Low=playing recordings less than 30 times; High=playing recordings greater than or equal to 30 times.

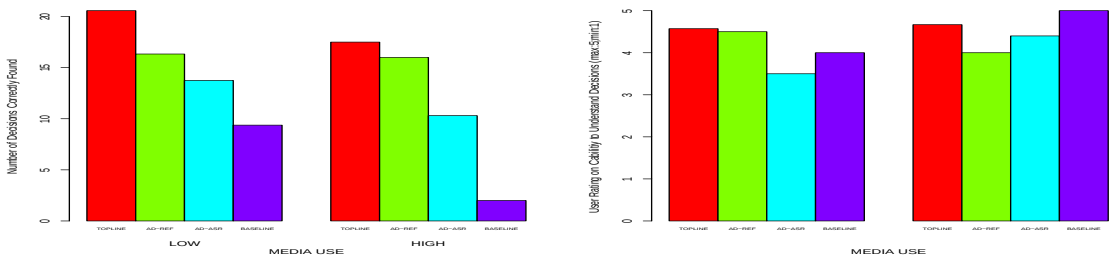

Figure 6: *Task effectiveness (number of decisions hits) and perceived success (user ratings on understanding all decisions) as a function of media usage.*

Fig. 6 demonstrates the effect of audio-video usage on task effectiveness and user-perceived success. The analysis reveals that the AD-ASR and Baseline users who turned to the audio-video browsing strategy (i.e. those with high usage of audio-video aids) were more likely to miss decisions in the archives. Interestingly, the lower task success rates did not affect the ratings of user-perceived success. For example, the group of high media usage users under the AD-ASR condition, who on average yielded lower task effectiveness, still perceived a high level of task success. The finding coincides with the subjects' comments that, although the audio-video recordings are difficult to use, they have provided grounds for decision understanding.

8.6 Conclusions

This study has verified our experimental hypothesis: Existing meeting summarization systems, which provide a general-purpose summary display, can be improved by refocusing the summaries with regard to user's information need. For users who require a quick overview of decisions, the decision-focused summary display was found to improve not only the actual task effectiveness, but also the overall report quality. Users also found the decision-focused summaries useful in helping them to achieve the task more effectively, e.g., finding all relevant information and understanding the decisions more efficiently. The browser interfaces that come with the decision-focused summary display are also rated as easier to use.

In addition, we evaluated the impacts of automation on the decision debriefing task. The findings are as follows: (1) The automatically generated decision-focused summaries, which contain 30%-40% inconsistencies with the gold standard manual summaries (as an average annotator), still assist users in producing high quality decision minutes and feeling confident about their performance on the decision debriefing task, despite some reduction in decision hits. (2) The ASR transcription has a greater negative impact on the actual task effectiveness and the quality of minutes. Another side effect of the ASR display is an increase in the level of user-perceived pressure.

Further investigation demonstrates a correlation between task effectiveness and usage of the summary display. As the content in the decision-focused summary is more closely tied to the user needs, participants who use these summaries (as opposed to the general purpose ones) rely more on the summary to find relevant information and, in turn, achieve higher performance.

Finally, the examination of user's media usage and coping strategies suggest that there exists an individual difference in the user's preference of whether to use the summary display or the audio-video playback facility to find relevant information. However, when the decision-focused summary is displayed with the ASR transcripts, users are often forced to view the audio-video recordings, since it is too difficult to use the other two displays. This hindered the performance of users who do not prefer to playback the audio-video recordings. This also suggests that there is a need to provide additional interface assistance to facilitate this group of users when ASR transcripts are used and may affect the comprehensibility of a succinct decision-focused summary. Possible interface enhances include a switching device that allows users to freely go from the view of a decision-focused summary back to that of a general-purpose summary.

References

AMIDA (2008). Amida proof-of-concept system architecture. Deliverable 6.7, AMIDA Integrated Project IST033812 (Augmented Multi-party Interaction with Distance Access).

Banerjee, S., Rose, C., and Rudnicky, A. I. (2005). The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In *Proceedings of the International Conference on Human-Computer Interaction*.

Barzilay, R. (2003). *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*. PhD thesis, Columbia University.

Carletta, J. (2006). Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. In *Proc. of LREC 2006, Genoa, Italy*, pages 181–190.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2006). The AMI Meeting Corpus: A pre-announcement. In Renals, S. and Bengio, S., editors, *Machine Learning for Multimodal Interaction II*, LNCS 3869, pages 28–39. Springer-Verlag, Berlin/Heidelberg.

Cremers, A. H., Hilhorst, B., and Vermeeren, A. P. (2005). What was discussed by whom, how, when and where? personalized browsing of annotated multimedia meeting recordings. In *Proceedings of HCI*, pages 1–10.

Edmundson, H. P. (1968). New methods in automatic extracting. *Journal of the ACM*, 16(2):264–285.

Hirschberg, J., Bacchiani, M., Hindle, D., Isenhour, P., Rosenberg, A., Stark, L., Stead, L., Whittaker, S., and Zamchick, G. (2001). SCANMail: Browsing and searching speech data by content. In *Proc. of Interspeech 2001, Aalborg, Denmark*, pages 1299–1302.

Hsueh, P.-Y., Kilgour, J., Carletta, J., Moore, J., and Renals, S. (2007). Automatic decision detection in meeting speech. In *Proc. of MLMI 2007, Brno, Czech Republic*.

Kleinbauer, T., Becker, S., and Becker, T. (2007). Combining multiple information layers for the automatic generation of indicative meeting abstracts. In *Proc. of ENLG 2007, Dagstuhl, Germany*.

Koumpis, K. and Renals, S. (2001). The role of prosody in a voicemail summarization system. In *Proc. of ISCA Workshop on Prosody in Speech Recognition and Understanding, Red Bank, NJ, USA*, pages 87–92.

Kupiec, J., Pederson, J., and Chen, F. (1995). A trainable document summarizer. In *Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA*, pages 68–73.

Lisowska, A., Popescu-Belis, A., and Armstrong, S. (2004). User query analysis for the specification and evaluation of a dialogue processing and retrieval system. In *Proceedings of LREC*, pages 993–996.

Mani, I. and Bloedorn, E. (1998). Machine learning of generic and user-focused summarization. In *Proceedings of the Fifteenth National Conference on AI (AAAI)*, pages 821–826.

Maskey, S. and Hirschberg, J. (2005). Comparing lexial, acoustic/prosodic, discourse and structural features for speech summarization. In *Proc. of Interspeech 2005, Lisbon, Portugal*, pages 621–624.

Matena, L., Jaimes, A., and Popescu-Belis, A. (2008). Graphical representation of meetings on mobile devices. In *MobileHCI 2008 Demonstrations (10th ACM International Conference on Human-Computer Interaction with Mobile Devices and Services)*, pages 503–506, Amsterdam.

Miekes, M., Müller, C., and Strube, M. (2007). Improving extractive dialogue summarization by utilizing human feedback‘. In *Proceedings of AIAP*, pages 627–632.

Murray, G. and Renals, S. (2007). Term-weighting for summarization of multi-party spoken dialogues. In *Proc. of MLMI 2007, Brno, Czech Republic*, pages 155–166.

Murray, G., Renals, S., and Carletta, J. (2005). Extractive summarization of meeting recordings. In *Proc. of Interspeech 2005, Lisbon, Portugal*, pages 593–596.

Pallotta, V., Niekrasz, J., and Purver, M. (2005). Collaborative and argumentative models of meeting discussions. In *Workshop on Computational Models of Natural Arguments (CMNA) at the IJCAI*.

Pallotta, V., Seretan, V., and Ailomaa, M. (2007). User requirements analysis for meeting information retrieval based on query elicitation. In *Proceedings of ACL*.

Popescu-Belis, A., Boertjes, E., Kilgour, J., Poller, P., Castronovo, S., Wilson, T., Jaimes, A., and Carletta, J. (2008). The amida automatic content linking device: Just-in-time document retrieval in meetings. In Popescu-Belis, A. and Stiefelhagen, R., 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.

Post, W. M., Cremers, A. H., and Henkemans, O. B. (2004). A research environment for meeting behavior. In *Proceedings of the 3rd Workshop on Social Intelligence Design*.

Rienks, R., Heylen, D., and van der Weijden, E. (2005). Argument diagramming of meeting conversations. In *Multimodal Multiparty Meeting Processing Workshop at the ICMI*.

Romano, N. C. and Nunamaker, J. F. (2001). Meeting analysis: Findings from research and practice. In *Proceedings of HICSS-34*. IEEE Computer Society.

Teufel, S. and Moens, M. (2002). Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Wellner, P., Flynn, M., Tucker, S., and Whittaker, S. (2005). A meeting browser evaluation test. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems 2005, Portland, OR, USA*, pages 2021–2024, New York, NY, USA. ACM Press.

Whittaker, S., Hirschberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., Stead, L., Zamchick, G., and Rosenberg, A. (2002). Scanmail: a voicemail interface that makes speech browsable, readable and searchable. In *Proc. of the SIGCHI 2002, Minneapolis, Minnesota, USA*, pages 275–282, New York, NY, USA. ACM.

Whittaker, S., Tucker, S., Swampillai, K., and Laban, R. (to appear). Design and evaluation of systems to support interaction capture and retrieval. *Personal and Ubiquitous Computing*.

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