FP6-506811

AMI

Augmented Multiparty Interaction

Integrated Project
Information Society Technologies

D2.4 Final Report on WP2 Activity

Due date: 31/12/2006
Project start date: 1/1/2004
LEAD CONTRACTOR: University of Edinburgh

Submission date: 31/12/2006
Duration: 36 months
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)	

D2.4 Final Report on WP2 Activity

Editor : Mike Lincoln (UEDIN)

Abstract:

This report summarizes and collates the work of WP2 - Infrastructure and Data Management, in AMI. The report includes a brief descriptions of the AMI Meeting Corpus including a description of the structure of the hub data, details of the AMI instrumented meeting rooms and information concerning the distribution mechanism for the corpus. We then go on to describe the collection of various spoke corpora including meetings using the AMI meeting browser and remote participation. Finally we describe background work supporting remote meetings for other AMI work packages.

1 The AMI Hub Corpus Overview

Any study of naturally-occurring behaviour such as meetings immediately encounters a well-known methodological problem: if one simply observes behaviour "in the wild", one's results will be difficult to generalize, since not enough will be known about what is causing the individual (or individuals) to produce the behaviour. For instance, if the group members have different roles or skills that bear on the task in different ways, that can naturally increase the importance for some contributions, and it can also be a deciding factor in whether the group actually needs to communicate at all or can leave one person to do all of the work. Vary any of the factors influencing the individuals behaviour and the data will change in character, but using observational techniques, it is difficult to get enough of a group history to tease out these effects. One response to this dilemma is not to make completely natural observations, but to standardize the data as much as possible by eliciting it in a controlled manner for which as many as possible of the factors are known. This approach, well-established in psychology and familiar from some existing corpora (e.g., [1]), comes with its own danger: results obtained in the laboratory will not necessarily occur outside it, since people may simply behave differently when performing an artificial task than they do in their daily lives.

Our response to this methodological difficulty is to collect our data set in parts. The first consists of elicited material using a design task in which the factors that affect individual behaviour are all fixed as far as they can be. The second consists of other, less controlled elicitations for different tasks. For instance, in one set of five meetings, forming one coherent set, which draws personnel from an existing work group to plan where to place people, equipment, and furniture in a fictionalized move to a new site that simplifies a real situation the group faces. The third contains naturally occurring meetings in a variety of types, the purpose of which is to help us validate our findings from the elicitation and determine how well they generalize by seeing how badly variation in the factors influencing participant behaviour affects our models.

Having decided upon the style of meetings to be included in the database, we next decided what data to record within the meetings. AMI researchers represent a diverse group with wide and varied research interests. The data collected must serve all these research areas and as such it was decided to make the captured data as all encompassing as possible. That is, as far as practical, anything and everything which occurs within the meetings is recorded. For each meeting, the following data is collected :

- Audio recordings, including far field recordings from microphones placed around the room, plus recordings from close talking microphones for each participant.
- Video recordings, including wide angle views of the entire meeting room, plus close up views of each participant.
- Images of any slides displayed on the projector.

- Recordings of any pen strokes made by participants on the whiteboard.
- Recordings of any handwritten notes made by participants.

All these recordings must be synchronised to allow events within the meeting to be cross referenced across modalities. Instrumented meeting rooms capable of the synchronised capture of this data have been developed at a number of partner sites.

Distributing the recorded meetings presented a further problem - The recordings themselves represent a huge quantity of data, and as such traditional means of data distribution (cd/dvd) are not feasible. To allow the data to be distributed between project partners, a central MultiModal Media File Server (the MMM server) was developed to allow partners access to the data as it becomes available. The MMM server is accessed via a web interface and contains media, annotations and documentation for the corpus.

1.1 The Scenario

In our meeting elicitation scenario [2], the participants play the roles of employees in an electronics company that decides to develop a new type of television remote control because the ones found in the market are not user friendly, as well as being unattractive and old-fashioned. The participants are told they are joining a design team whose task, over a day of individual work and group meetings, is to develop a prototype of the new remote control. Within this context, each participant in the elicitation is given a different role to play. The *project manager* (PM) coordinates the project and is responsible overall. The *marketing expert* (ME) is responsible for determining user requirements, watching market trends, and evaluating the prototype. The *user interface designer* (UI) is responsible for the technical functions the remote control provides and the user interface. Finally, the *industrial designer* (ID) is responsible for designing how the remote control works including the componentry. For this elicitation, we use participants who are neither professionally trained for design work nor experienced in their role. While it is well-known that expert designers behave differently from novices, using professional designers for our collection would present both economic and logistical difficulties. Moreover, since participants will be affected by their past experience, all those playing the same role should have the same starting point if we are to produce replicable behaviour. To enable the participants to carry out their work while lacking knowledge and experience, they are given training for their roles at the beginning of the task, and are each assigned a (simulated) personal coach who gives sufficient hints by e-mail on how to do their job.

1.1.1 Structure of the Scenario

[3] distinguishes the following four phases in the design process: Project kick-off; Functional design; Conceptual design; Detailed design. We use these phases to structure our elicitation, with one meeting per design phase. In real groups, meetings occur in a cycle

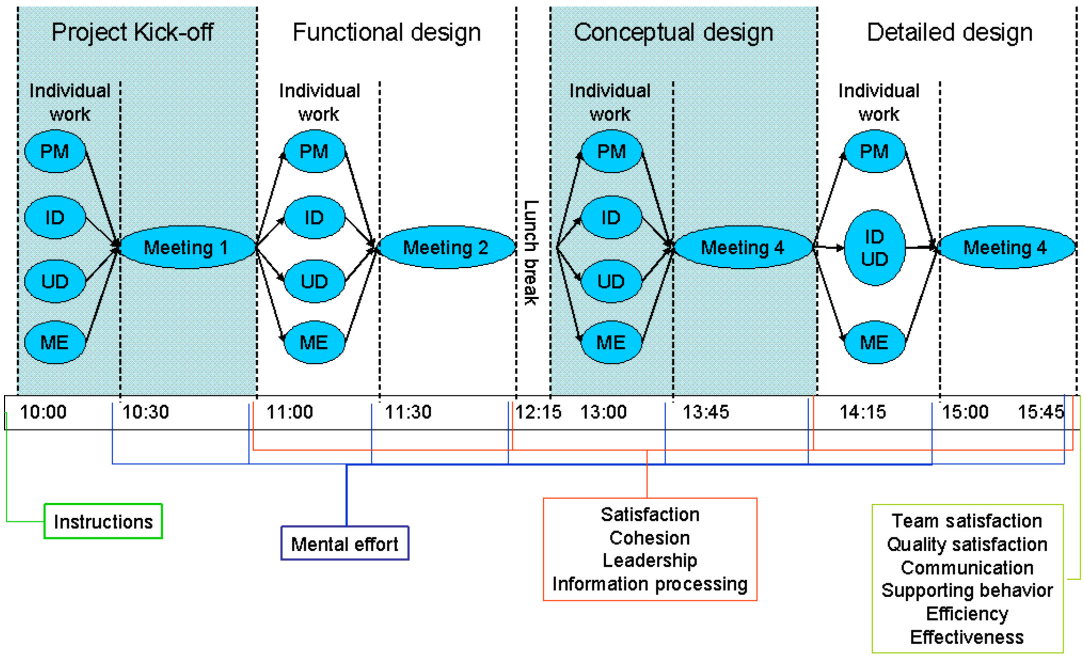


Figure 1: The meeting paradigm: time schedule with activities of participants on top and the variables measured below. PM: Project Manager; ID: industrial designer; UI: user interface designer; ME: marketing expert.

where each meeting is typically followed by production and distribution of minutes, the execution of actions that have been agreed on, and the preparation of the next meeting. Our groups are the same, except that for practical reasons, each design project was carried out in one day rather than over the usual more extended period, and we included questionnaires that will allow us to measure process and outcomes throughout the day. An overview of the group activities and the measurements used is presented in fig. 1.

1.1.2 The Participants Environment

The setting for the scenario is as close as possible to a traditional office environment. The subjects each have a private office, and a personal laptop with Microsoft Office tools, e-mail, and a web browser. During the meetings, which are held in the instrumented meeting rooms, the participants are provided with typical meeting room apparatus such as a whiteboard and overhead projector. Throughout the day, information and instructions are provided to each of the participants by means of their laptop computers from a specially developed ‘scenario tool’ as shown in figure 2. Information is provided in one of three forms : *Emails* provide information and tutoring concerning the participants role. *Pop-up messages* give instructions or warnings to the participants which require immediate action. *Web pages* are accessible form a restricted ’simulated web’.

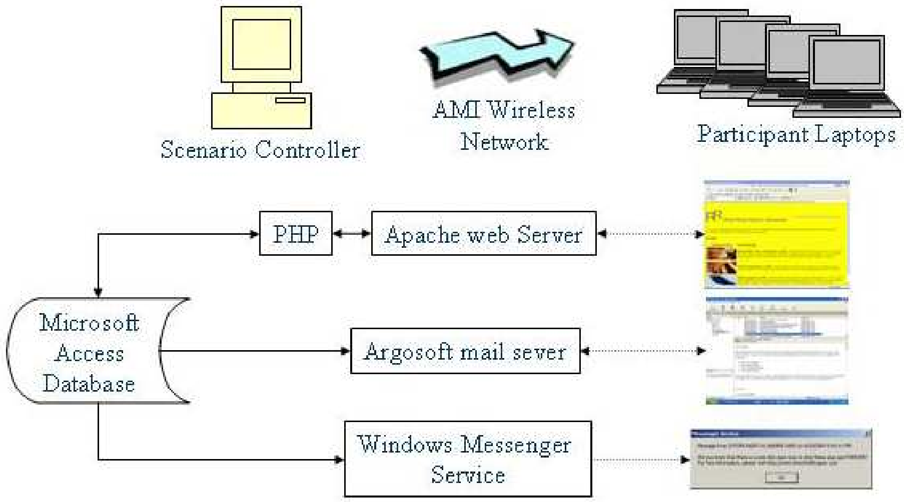


Figure 2: Overview of the Scenario Controller Architecture.

The controller itself is an MS Access database application, where each entry in the database is an event. An event consists of the information provided (E.G. a web page, an email, a popup message), the means by which it is provided (I.E. one of 'email', 'popup' or 'web'), the time the event occurs, and the participant to whom the information is delivered. At the initiation of the project, the database is launched and it schedules the list of events, which are then processed as the project runs. By providing information and instructions at specific times the participants are encouraged to follow the project method described above while still behaving in a natural manner.

1.2 The AMI Instrumented Meeting Rooms

Instrumented meeting rooms have been constructed at The University of Edinburgh (UEDIN), IDIAP, and TNO Human Factors (TNO) for the collection of the database. In the following sections, the Edinburgh meeting room and the data collected will be described in detail, and differences between this and the other rooms will subsequently be outlined.

1.2.1 The UEDIN Room

An overhead schematic view of the UEDIN meeting room is given in figure 3, and a diagram showing the connectivity of the audio/visual capture equipment is shown in figure 4.

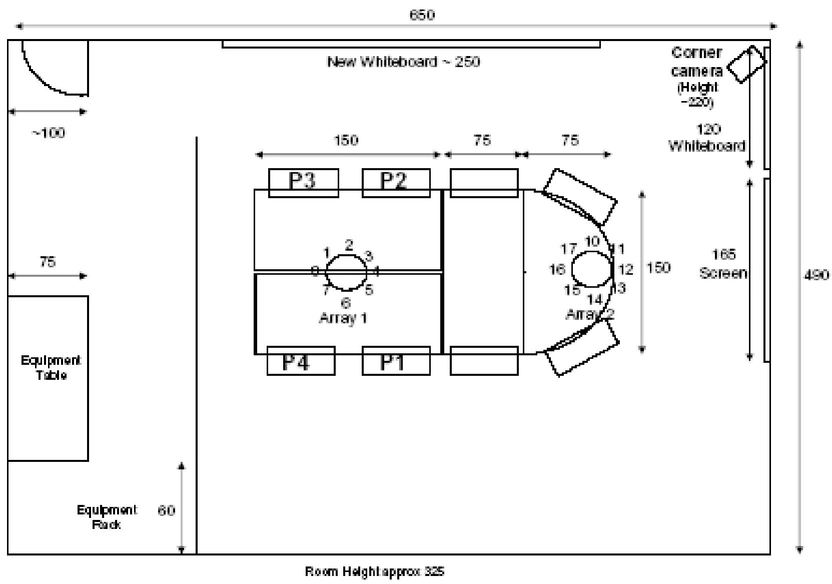


Figure 3: Overhead Schematic View of the UEDIN Instrumented Meeting Room, Showing the Participant Positions (P1-4), the Microphone Arrays, and the Wide Angle Camera Position.

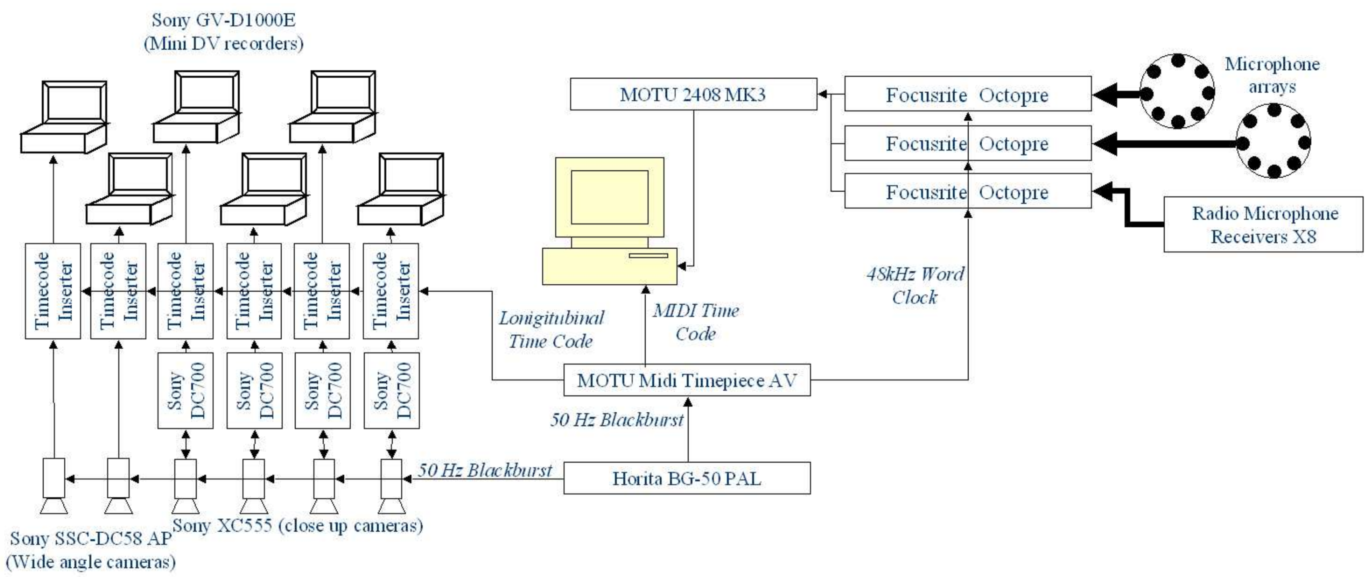


Figure 4: Diagram of the connectivity of the A/V Capture Equipment in the UEDIN room.

A. Closeup | B. Corner | C. Overhead

Figure 5: Camera views from the UEDIN Instrumented Meeting room.

Audio

The room contains 24 microphones in total. 16 omni-directional microphones are arranged in two 10cm radius circular arrays of 8. In addition each of the four participants wears a radio lapel and radio headset microphone. Using a radio based system allows participants the same freedom of movement they would have if not wearing microphones, while providing audio of the same quality as wired mics. Three 8 channel microphone pre-amplifiers are used to amplify and digitize the microphone outputs at 48KHz, 16 bit resolution. A 24 channel PCI audio interface is then used to transfer the audio data to a pc where it is recorded directly to disk using standard audio recording software.

Video

Video is recorded straight to mini-dv tape from 6 cameras in the room. 4 subminiature cameras mounted under the central microphone array provide close up views of each of the participants as shown in figure 5A. 2 CCTV cameras provide wide angle views of the room - one mounted above the center of the table gives an overhead view of the entire floor area of the room, while a second mounted in the corner of the room provides a view of the whiteboard and presentation areas (figures 5B and C).

Synchronization

Without further hardware to provide synchronization signals, the following errors would occur during data acquisition: The audio A to D converters would sample each channel at different times; The cameras would acquire frames at different times; With no global timestamp recorded on all streams, it would be impossible to accurately re-merge the individual channels for subsequent processing.

To alleviate these problems we use the following synchronisation technique : A Blackburst Generator provides a composite video timing signal which is used as a reference signal to which all other devices are locked. The signal is fed directly to each of the video cameras to ensure they sample frames at exactly the same instant. A further output is connected to a MIDI Timepiece AV unit which generates the following timing signals from the blackburst reference : *48 kHz word clock*. This is used to trigger the audio A to D converters ensuring that each channel is sampled at precisely the same instant. *Longitudinal Time Code (LTC)*. This is Hours:Minutes:Seconds:Frames time code encoded as an 80bit word for each video frame, and subsequently inserted in each video frame as a series of black and white dots. *MIDI Time Code*. This is the LTC output in a format which can be read by MIDI devices. In the meeting room it is read by the recording software and used to timestamp the audio samples. Because the same LTC is used to timestamp the audio and video signals, they can be accurately synchronised as a post processing step.

Auxiliary Data In addition to audio and video, a second PC is used to capture the auxiliary data. The same MIDI Time Code is used to timestamp the data and ensure it is synchronized with the audio and video streams. *Whiteboard*: A digital white board system is used to capture any pen strokes the participants make on the whiteboard. *Beamer*: Any slides presented on the beamer are captured via a VGA capture card and stored as jpeg images. *Handwritten notes*: Each participant has access to a digital pen throughout the scenario. The pen stores the time stamped x-y co-ordinates of any pen strokes made on special paper which are then downloaded to the Auxiliary Capture PC as xml files for subsequent processing.

1.2.2 Deviations of other rooms

In addition to being installed in rooms with significantly different acoustic properties, the TNO and IDIAP rooms have some small differences in their recording equipment:

TNO

The TNO room contains a 10 element linear array in place of one of the circular arrays; Each participant has a headset mounted radio microphone, but no lapel mic; The position of the wide angle cameras is different; Three windows XP computers, each fitted with video capture cards are used to capture and encode the video data and stream it directly to hard disk, rather than recording to digital video tapes.

IDIAP

The IDIAP room has 3 wide angle video cameras rather than the two in the UEDIN room. The second IDIAP circular microphone array has only 4 elements; A binaural manikin is placed at the end of the table furthest from the screen providing 2 further audio channels.

1.3 Data pre-processing

Before it can be made available to data consumers, some pre-processing of the data is required, e.g. to encode the digital video tapes, and to ensure all the media channels are properly synchronised.

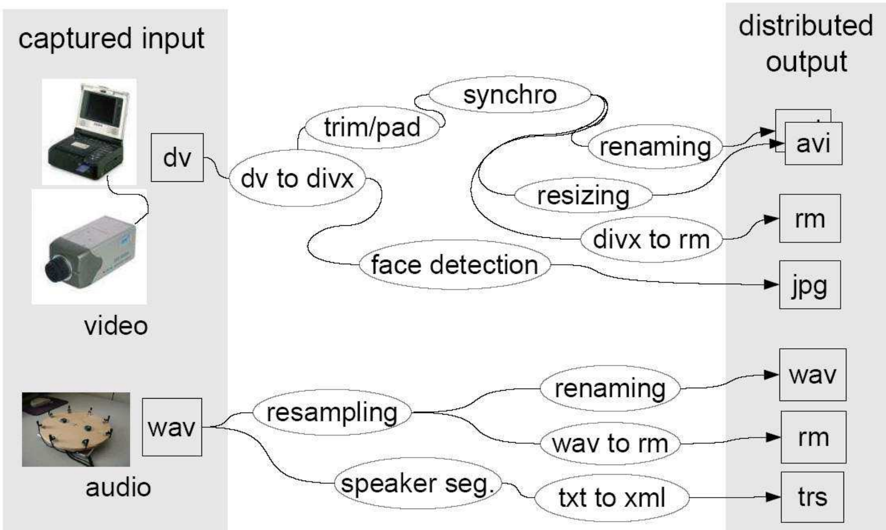


Figure 6: Audio and Video Pre-Processing

1.3.1 Audio and Video Pre-Processing

Figure 6 shows the pre-processing stages for the audio and video data.

Video In the case of data recorded to tape, the digital video tapes are digitised and stored to disk as full frame DivX AVI files. Quarter frame videos are also made available to save storage space and download times for users who do not require full frame resolution. Realmedia files of the combined audio and video are also produced and are available for download, or streaming.

Audio: The audio is downsampled from 48kHz to 16kHz, and made available as a wav files - 24 for each meeting (one for each audio channel). The data is also encoded in realmedia format for streaming.

To aid in the transcription of the meetings,a speech/silence segmentation for each person in a meeting is made from their lapel microphone recordings.

1.3.2 Auxilliary Data Pre-Processing

Figure 7 shows the pre-processing stages for the Auxilliary data

The XML files generated by the E-Beam whiteboard capture system and the Logitech IO pens are converted into jpeg images showing what the participants wrote. A DIV-X movie of these files, synchronised to the audio and video recordings is also produced. A transcription of the written data is also generated using an optical character recognition technique based on [4].

Optical character recognition based on the technique described in [5] is used to produce transcriptions of the captured slides. This, along with the captured jpeg images and

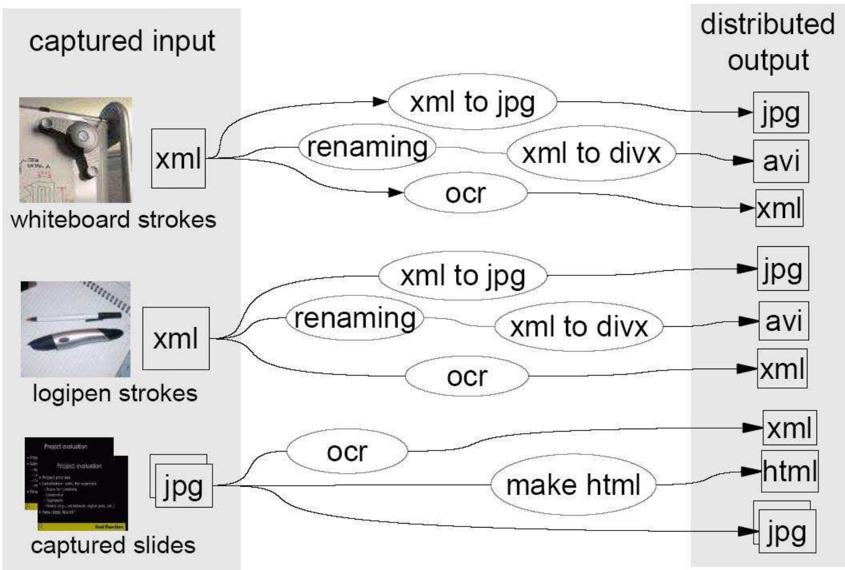


Figure 7: Auxiliary Data Pre-Processing Stages

html pages containing the images is made available as part of the database.

1.4 Data Distribution

All data included in the AMI Meeting Corpus, with the exception of the full frame video files, is stored on and distributed online from the IDIAP MultiModal Media fileserver (MMM server). The latest version of the MMM server, which currently has space for 3Tb of data, is based on Plone CMS which offers a unified framework to share project data, and a more powerful and flexible management of individual access rights to the data than the initial system. Access to the Hub corpus is via the AMI Meeting Corpus website (http://corpus.amiproject.org). The site contains full documentation for the corpus and includes an interactive web interface based on CGI dynamic pages providing on-demand data packaging. This allows users to download any combination of media and annotations from the corpus they require. While the single download site at IDIAP has proved sufficient to meet the download demands for the corpus, the data is mirrored at 2 additional sites - UEDIN and BRNO, which, as well as providing additional backups of the data, can easily be used to provide additional download capacity if demand requires. Because of the quantities of data involved it is impractical to download the full frame videos over the internet. Users who require this data may purchase it (at cost) on firewire hard disks. To complement the web site, a “taster” DVD has been produced containing media and annotation samples, example browsers and sample annotation tools. The DVD has been widely distributed at conferences and exhibitions to interested parties.

1.5 Corpus contents

A total of approximately 100 hours of data are included in the hub corpus. This data consist of 72 hours of design scenario data, 5 hours of less controlled elicitations (as described in 1) 23 hours of 'real' naturally occuring meetings.

Design scenario meetings. Of the 72 hours of scenario meetings, 30 hours have been recorded in the UEDIN room (meetng IDs ES20*), 18 hours have been recorded in the IDIAP room (meeting IDs IS10*) and 24 hours have been recorded in the TNO room (meeting IDs TS30*)

Other elicited meetings These meetings constitute 5 hours of recordings made in the IDIAP room (meeting IDs IB40*). They consist of recordings of meetings concerning the move of a working group to new office space, and recordings of the meetings of a book club.

'Real' meetings These meetings contain 23 hours of recordings made in the UEDIN and IDIAP rooms (meeting IDs EN* and IN*). They consist of meetings that were happening anyway, such as meetings of MSc students involved in a group design project and meetings associated with research projects.

2 Spoke corpora

In addition to the Hub recordings, a number of spoke copora have been collected which either provide data for a specific AV procesing tasks, have been recorded during the evaluation of AMI technologies, or have been collected to obtain feedback from project partners concerning the addition of new technology to the meeting rooms.

2.1 Mobile meeting room data

The Brno university of technology (BUT) have developed a low-cost, mobile, meeting capture system built from off-the-shelve products which is easy to install and operate. The meeting room incorporates a 360 degree image capturing system using a hyperbolic mirror and subsequent image-processing.

2.1.1 Capture system

Audio from the participants is captured from 4 lapel microphones and recorded directly to disk on a laptop computer. Video is captured from an HDTV camera directed at the hyperbolic mirror and again captured directly to the laptop computer. The camera, and the view from the hyperbolic mirror is shown in figure 9. The camera also provides two further microphones, capturing audio from the entire meeting environment. The total cost of the hardware is approximately 4000 Eur. excluding the computer, and is easily

Figure 8: Mobile meeting room.

transported by a single person. The capture software was developed at BUT and provides synchronised capture of all audio and video.

2.1.2 Post processing

Due to the construction of the meeting room hardware the camera is susceptible to vibrations which would cause undesirable movement in the final unwrapped image. To overcome this a realtime system, based on [6], has been developed which provides image stablisation at a sub-pixel level. The image is then unwrapped to form a panoramic image as shown in (Fig. 10)

Figure 9: HDTV camera and the original image.

Figure 10: Unwrapped image from hyperbolic mirror.

2.1.3 Recording status

20 hours of czech-english meetings have been captured using the mobile meeting room and are available to AMI partners.These data were mainly used for work on video tracking and automatic video editing algorithms [6, 7].

2.2 AV16.7.ami

This corpus consists of 16 sequences recorded from two camera angles (mean duration of 99 seconds). Seven sequences were designated as the training set, and nine sequences for testing. The sequences depict up to four people in a meeting-room performing common actions such as sitting down, discussing around a table, etc. Participants acted according to a predefined agenda (they were told the order in which to enter the room, sit, or pass each other), but the behavior of the subjects was otherwise natural. The sequences contain many challenging phenomena for tracking methods including occlusion, cameras blocked by passing people, partial views of backs of heads, and large variations in head size. The corpus has also been annotated using bounding boxes for head location for use in training and evaluation.

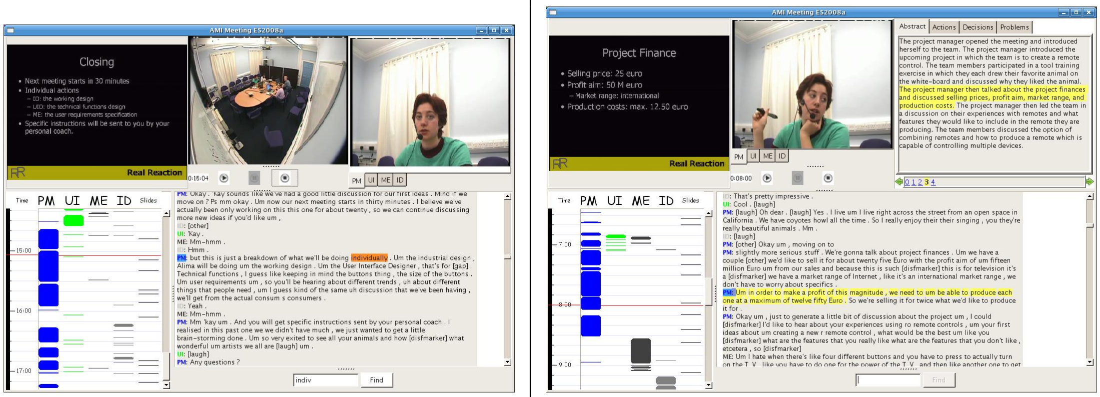
2.3 The Multi-Channel Wall Street Journal Audio Visual corpus (MC-WSJ-AV)

The MC-WSJ-AV corpus offers an intermediate task between simple digit recognition and large vocabulary un-constrained conversational speech for speech recognition in meeting environments. The corpus consists of read Wall Street Journal sentences taken from the test set of the WSJCAM0 database, recorded in the AMI instrumented meeting rooms. The sentences are read by a range of speakers (30 in total) with varying accents (including a number of non-native English speakers). Sentences are read according to a number of scenarios including a single stationary speaker, a single moving speaker, and multiple concurrent speakers. The full AMI data capture system is used during recordings, providing lapel, headset and microphone array audio data from each participant and synchronised close up and wide angle video views of the speakers. The data is suitable for a wide variety of research tasks including : development of microphone array ASR frontend processing systems; audio-visual ASR; audio-visual person tracking; integration of audio-visual person tracking with microphone array ASR processing; recognition of accented and non-native English speech; recognition of overlapped speech. Data from the concurrent speakers section of MC-WSJ-AV is currently being in the PASCAL Speech Separation Challenge II.

2.4 The Task Based Evaluation

The Task Based Evaluation (TBE) is a data collection effort aimed at evaluating the meeting browser technology developed by the AMI project. The basic methodology consists of inviting groups of 4 participants to conduct only the *forth* meeting in the AMI remote control design scenario. In order to prepare for the meeting the new groups are provided with recordings of the first three meetings from a previously recorded group taken from the hub corpus. The new group must use the recordings from the previous group to gather information about the project and the decisions they came to, before conducting the final meeting. The original AMI scenario controller was updated, and new infrastructure added to the IMR to allow the new groups to view the previous recordings in one of three ways: Condition zero is simply a directory containing the audio and video recordings of the meetings, along with the documents the group produced. Condition one is a meeting browser with synchronised audio and video from each participant, presentation slides, speaker activity display and searchable speech transcription as shown in figure 11, left. Finally condition two provides a browser with the facilities of condition one, plus higher level information such as abstracts, action points, and decisions for each meeting, indexed back to the relevant position in the media 11, right. Ten groups of participants performed each of the three conditions.

For each condition five groups were recorded in the UEDIN Instrumented meeting room, and five at TNO. Recordings were not conducted at IDIAP due to the problem of locating English speaking participants who had not been involved in the original recordings of the AMI hub data. At the UEDIN site, participants were recorded using identical equipment to that used for the original hub recordings. In addition, to allow analysis of the use of the browsers, Camtasia screen capture software recorded the participants interactions with their laptops. At TNO a lower quality capture system was used since the use of cheaper, lower quality sensors is one of the focuses for the AMIDA project. Here, two laptop computers run the Quindi meeting companion system, with each machine capturing video of 2 participants using a web cam and audio from a desktop boundary microphone. Again the participants screens are captured to allow subsequent analysis of their use of the browsers. Data from the UEDIN recordings has been supplied to NIST and is being transcribed for use in the RT07s rich transcription evaluations.



Condition one | Condition two

Figure 11: Screen shots from the condition one (left) and condition two (right) TBE evaluation meeting browsers.

2.5 Remote Participation Meetings

Trial meetings including remote participation via the Visual Nexus video conferencing system have been recorded. The trial followed the same procedure as the condition zero TBE recordings, except that one of the participants was located in a separate office and joined the meeting being conducted in the instrumented meeting room remotely. In the instrumented meeting room a digital video camera provides a view of the meeting room table for the video conference, and audio is taken form a single boundary microphone placed on the meeting room table. The Particpant playing the “Project Manager” role controls the visual nexus system, including displaying the powerpoint presentations via Visual Nexus’ application sharing system, with the conference being displayed to other partipants using the meeting room projector.

The “remote” participant uses a low cost web camera with built in microphone to join the conference, and Visual Nexus’s conference recording capabilities are used to record the conference from these low cost sensors. The proximity of the remote participants office to the IMR meant that the standard radio microphones used in previous recordings were also used to provide high quality audio capture of their speech. This will allow the data to be used to compare performance of audio processing algorithms on data captured with both high and low quality sensors. High quality video from the remote participant is provided from a digital video camera located within their office. The meetings are currently being assessed by AMI partners to determine the recording system’s suitability for recording further AMIDA meetings.

3 Supporting tasks

3.1 Enhancement of Visual Nexus systems and support for remote project meetings

A Visual Nexus online multimedia meeting service was made available to the AMI community to allow the work groups to experience remote attendance of meetings and in particular to begin evaluation of the concepts of a remote meeting assistant. The first group of AMI members was connected at the end of 2005, the majority given access in May 2006. The endpoint included the ability to record the audio/video calls and meetings. At the end of 2006 a utility to convert the recordings from proprietary to windows media format was created and made available for use in January 2007.

In the second half of 2006, Visual Nexus and AMI partner RealVNC began to collaborate on the remote control program tool, called VNC: (a) to extend VNC into an application shareign tool allowing Visual nexus users to share individual applications as opposed to the whole desktop. (b) to enhance VNC to work in multi-party meeting mode as well as peer-to-peer mode. (c) To integrate the enhanced VNC into the Visual Nexus online meeting system, so that VNC could be utilised to share applications in the context of remote online multi-media meetings. Specification of the work has been completed and development of the prototype is well underway. Visual Nexus participation during 2006 has been at the specification and architectural level. The development work to integrate VNC with Visual Nexus is under way but not yet complete.

3.2 Development of Realtime Annotation Distribution System - The Hub

For real time meeting applications there is a strong need for a robust real time annotation data distribution mechanism and a prototype of such a system, called Hub, has been implemented. It is designed to handle both past archived meetings as well as live meeting processing and browsing. As an example, this will allow online audio and visual-based recognizers to be able to add data to it and to have multiple interested clients (ie. JFerret meeting browser) to receive new information in real time. Hub includes a flexible model for the conversion of NITE annotations to a database format suitable for the JFerret meeting browser and allows multiple clients to simultaneously produce or consume annotations from the centralized server. The model has been tested on the full AMI annotation corpus, the conversion being validated by re-generating the NITE annotations from the annotation database.

References

[1] Anderson, A.H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., Weinert, R.: The HCRC Map Task Corpus. Language and Speech **34** (1991) 351–366

[2] Post, W.M., Cremers, A.H., Henkemans, O.B.: A research environment for meeting behavior. In Nijholt, A., Nishida, T., Fruchter, R., Rosenberg, D., eds.: Social Intelligence Design, University of Twente, Enschede, the Netherlands (2004)

[3] Pahl, G., Beitz, W.: Engineering design: a systematic approach. Springer, London (1996)

[4] Liwicki, M., Bunke, H.: Handwriting recognition of whiteboard notes. In Marcelli, A., ed.: 12th Conference of the International Graphonomics Society, Salerno (2005)

[5] Chen, D., Odobez, J.M., Bourlard, H.: Text detection and recognition in images and videos. Pattern Recognition (2004)

[6] Sumec, S., Potucek, I., Zemcik, P.: Automatic mobile meeting room. In: Proceedings of 3IA'2005 International Conference in Computer Graphics and Artificial Intelligence. (2005)

[7] Sumec, S.: Multi-camera automatic video editing. In: CCVG 2004, Warsaw, Poland. (2004)