



**FP6-506811**

**AMI**

**Augmented Multiparty Interaction**

Integrated Project  
Information Society Technologies

## **D2.2 The AMI Multimodal Meeting Database - Infrastructure, Data and Management**

**Due date:** 30/06/2005

**Submission date:** 07/08/2005

**Project start date:** 1/1/2004

**Duration:** 36 months

**Revision:** 1

**IDIAP**

<b>Project co-funded by the European Commission in the 6th Framework Programme (2002-2006)</b>		
<b>Dissemination Level</b>		
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)	



## D2.2 The AMI Multimodal Meeting Database - Infrastructure, Data and Management

Editor : Mike Lincoln (UEDIN)

**Abstract:** This report gives a detailed description of the multimodal meetings database recorded for the AMI project. The Hub database contains recordings of a variety of styles of meetings, some naturally occurring and some elicited, collected in instrumented meeting rooms at a number of participant sites. The report documents the data set, including the equipment used to obtain the elicited meetings; detailed descriptions of the instrumented meeting rooms, the equipment they contain and the captured data; subsequent processing of the raw data prior to distribution and the Media File Server used to distribute the data. We also include a description of a spoke data set recorded using a mobile meeting room.

# 1 Database 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. [1] identifies seven kinds of factors that affect how work groups behave, ranging from the means they have at their disposal, such as whether they have a way of communicating outside meetings, to aspects of organizational culture and what pressures the external environment places on the group. The type of task the group is trying to perform, and the particular roles and skills the group members bring to it, play a large part in determining what the group does; 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 these factors 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. Experimental control allows the researcher to find effects with much greater clarity and confidence than in observational work. This approach, well-established in psychology and familiar from some existing corpora (e.g., [2]), 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 [1] describe 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. These again provide more control than in natural data, but give us a first step towards thinking about how one combines data from disparate sources. 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 affects our models. The goal in this part of the collection was not to constrain the type of meeting in any way apart from keeping the recording manageable, but to allow the factors to vary freely. Taking histories that would allow us to classify the groups by factor would be a formidable task, and so the recorded data is included “as is”, without supplementary materials.

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 pen strokes made by participants on notepads supplied to them.

All these recordings must be synchronised to allow events within the meeting to be cross referenced across modalities. Instrumented meeting rooms capable of the 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 MultiMedia File Server (the MMM server) was developed to allow partners access to the data as it becomes available. In addition, a wiki is maintained with information concerning each meeting which may not be available from the captured data, E.G. participant information, details of any recording anomalies or details of any special post processing applied to the data

In the following sections we describe the elicitation scenario used to record the first section of the data, and how it is implemented by means of a ‘scenario controller’. We then go on to give a detailed description of the instrumented meeting rooms, the data which is captured, and how it is recorded and subsequently processed. We then describe the MMM server used to distribute the data and its interface, and finally we give a detailed break down of the size and composition of the database recorded and currently available on the MMM server.

## 2 The meeting elicitation scenario

In our meeting elicitation scenario [3], 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. We chose design teams for this study for several reasons. First, they have functional meetings with clear goals, so making it easier to measure effectiveness and efficiency. Second, design is highly relevant for society, since it is a common task in many industrial companies and has clear economic value. Finally, for all teams, meetings are not isolated events but just one part of the overall work cycle, but in design teams, the participants rely more heavily on information from previous meetings than in other types of teams, and so they produce richer possibilities for the browsing technology we are developing.

### 2.1 Participants and roles

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. His job is to guarantee that the project is carried out within time and budget limits. He runs the meetings, produces and distributes minutes, and produces a report at the end of the trial. 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. The user interface designer and industrial designer jointly have responsibility for the look-and-feel of the design.

For this elicitation, we use participants who are neither professionally trained for design work nor experienced in their role. It is well-known that expert designers behave differently from novices. However, 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

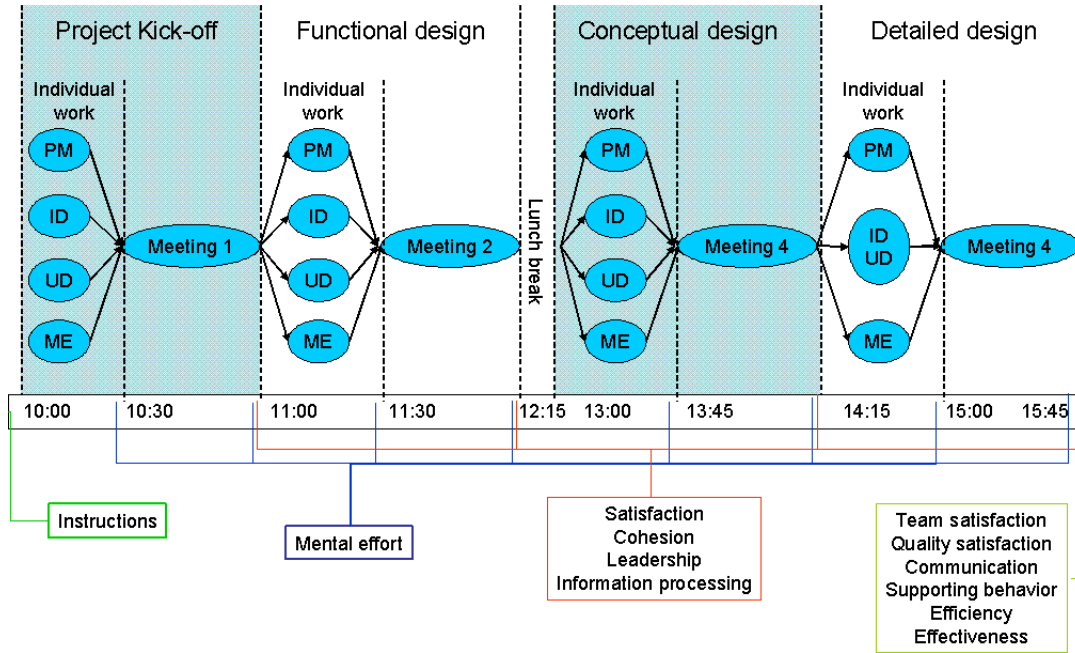


Figure 1: The meeting paradigm: time schedule with activities of participants on top and the variables measured below. PM: Project Manager; ID: industrial designer; UI: user interface designer; ME: marketing expert.

by e-mail on how to do their job. Our past experience with elicitations for similar non-trivial team tasks, such as for crisis management teams, suggests that this approach will yield results that generalize well to real groups. We intend to validate the approach for this data collection both by the comparisons to other data already described and by having parts of the data assessed by design professionals.

## 2.2 The structure of the elicited data

[4] distinguishes the following four phases in the design process:

- *Project kick-off*, consisting of building a project team and getting acquainted with both each other and the task.
- *Functional design*, in which the team sets the user requirements, the technical functionality, and the working design.
- *Conceptual design*, in which the team determines the conceptual specification for the components, properties, and materials to be used in the apparatus, as well as the user interface.
- *Detailed design*, which finalizes the look-and-feel and user interface, and during which the result is evaluated.

We use these phases to structure our elicitation, with one meeting per design phase. In real groups, meetings occur in a cycle 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. In future data collections we intend to collect further data in which the groups have access to meeting browsing technology, and these measures will allow us to evaluate how the technology affects what they do and their overall effectiveness and efficiency. An overview of the group activities and the measurements used is presented in fig. 1.

### 2.3 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 described in section 3, the participants can use a PowerPoint projector for presentations and an electronic whiteboard. Participants may also exchange files by use of a 'public folder' which is common to all laptops, and store information in their own 'private documents' folder.

### 2.4 Information provision

Throughout the day, information and instructions are provided to each of the participants by means of their laptop computers from a special developed 'scenario tool' as shown in figure 2. Information is provided in one of three forms :

1. **Email** : The controller sends e-mails in the name of (virtual) individuals outside the team to participants within the team. In this way, for instance, the 'account manager' can provide information on the design problem, and the 'personal coach' can bring in knowledge and experience about the participants role.
2. **Pop up message** : The Microsoft messenger service is used to send popup messages from the controller that gives instructions or warnings to participants. For example, one alert instructs the project manager to round up the meeting, and another alert warns all participants they have five minutes left before the next meeting starts.
3. **Web pages** : Participants do not have access to the entire internet. Instead, the scenario controller defines the availability of information the participants can find on a 'simulated' web. In this way, the marketing expert, for example, can only search and browse market trends that we choose to make available at a particular moment. The web server is also used to collect questionnaires which the participants are asked to fill in at specific times during the project which are then distributed as part of the database.

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. For example, in figure 3, one can see that 4 hours and 25 minutes after the beginning of the project event number 129 is launched and all participants receive an alert from the messenger server with the message '5 minutes to round up meeting preparations'.

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.

A complete set of the information sent to the participants is available at <http://www.amiproject.org/private/WP02/Docs>

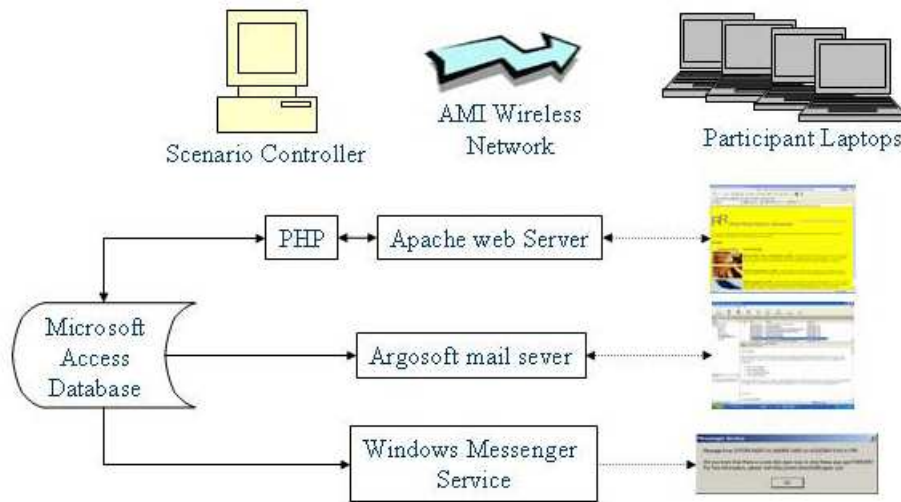


Figure 2: Overview of the Scenario Controller Architecture.

### 3 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.

#### 3.1 The UEDIN Room

An overhead schematic view of the UEDIN meeting room is given in figure 4, and a diagram showing the connectivity of the audio/visual capture equipment is shown in figure 5.

##### 3.1.1 Audio

24 mono audio channels are recorded directly to hard disk using the following equipment :

**Microphones.** The room contains 24 microphones in total. 16 Sennheiser MK2E-P-C miniature omni-directional electret microphones are arranged in two 10cm radius circular arrays of 8. These are placed in the center of the meeting room table, one between the participants and one at the end of the table closest to the presentation screen and whiteboard. The MK2E-P-C was chosen for its linear frequency response from 20Hz to 20kHz, and it's ability to draw phantom power directly from the microphone pre-amplifier. 8 Sennheiser EW300 Series radio microphones are used for recording the four participants. Each person wears an ME 3-N close talking headset condenser mic and an MKE 2-EW omni directional lapel mic, the later being a wireless equivalent to those used in the arrays. 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.

**Preamplifiers and Analogue to Digital (A to D) conversion.** Three Focusrite Octopre 8 channel microphone pre-amplifiers with up to 24 bit 96 kHz analogue to digital converters are used

The screenshot shows the 'A-M Scenario Controller - [FormEvent - Form]' window. The main area is a table with columns: EventID, Time, RoleAssignment, InfoAvailability, Hyperlink, MessageText, BodyFile, and Body. The table contains 19 rows of event data. Below the table is a 'Records' section showing '1 of 21' records. At the bottom is a 'Scenario time control' panel with fields for 'Elapsed time' (04:10:12), 'Start time' (09:00:00), 'End time' (18:00:00), 'Scenario time' (12:10:12), 'Accelerator' (1), 'Real time' (14:44:18), and 'Events processed' (120). A 'Scenario is running' indicator is present.

EventID	Time	RoleAssignment	InfoAvailability	Hyperlink	MessageText	BodyFile	Body
129	04:25:00	all			5 minutes to round up meeting preparations		
130	04:26:00	all			Please, check your email to fill in questionna		
131	04:26:00	all			Please fill in questionnaire 7	Questionnaire07Mail.html	
132	04:30:00	participant 1			Move to meeting room and bring your laptop		
133	04:30:00	participant 4			Move to meeting room and bring your laptop		
134	04:35:00						
135	04:35:00	participant 1					
136	04:35:00	participant 1					
137	04:35:00						
138	04:40:00	participant 2					
139	04:45:00	participant 3					
140	04:50:00	participant 4					
141	04:55:00	all					
142	05:10:00	participant 1			warning: 5 minutes to finish meeting		
143	05:15:00	all			warning: finish meeting now		
144	05:16:00	all			Please, check your email to fill in questionna		
145	05:16:00	all			Please fill in questionnaire 8	Questionnaire08Mail.html	
146	05:21:00	all			Please, check your email to fill in questionna		
147	05:21:00	all			Please fill in questionnaire 9	Questionnaire09Mail.html	
148	05:31:00	all			Please, check your email to fill in questionna		
149	05:31:00	all			Please fill in last questionnaire	QuestionnairePostMail.html	

Figure 3: Example entries in the Scenario Controller Database

to amplify and digitize the microphone outputs. Each channel has a separate class A amplifier with independent gain control, and digitized output is via a single ADAT Lightpipe fibre optic cable carrying all 8 channels. The A to D converters can sample at a variety of rates using either the Octopre's internal clock or from an external source via a word-clock input - Here the data is captured at 48kHz, 16bit resolution. The Octopres also provide phantom power for the MK2E-P-C microphones.

**Audio I/O.** The Mark of the Unicorn (MOTU) 2408 MKIII is an audio interface for PC based hard disk audio recording. It consists of a 19" rack mounted I/O unit connected via a Firewire like interface to a PCI card installed in the PC. The I/O unit supports 24 input/output channels in 3 banks of 8, with all 24 channels capable of operating simultaneously. Software installed on the PC allows configuration and acquisition of each of the channels via the PCI card. In the meeting room, each of the ADAT Lightpipe outputs from the Octopre's A to D converters are connected to one bank of a single I/O unit and are subsequently acquired by the PC via PCI card.

**Audio Capture Computer.** The audio capture computer is a 3GHz P4 with two 40MB SCSI hard drives configured as a RAID 0 array for streaming audio to. The operating system used is Windows XP for compatibility with the MOTU driver software, and audio is captured and exported using Cakewalk Sonar recording software.

### 3.1.2 Video

Six channels of video are recorded to mini-DV tape as described below :

**Cameras.** 6 cameras in total are used to record the video in the room. 4 Sony XC555 subminiature cameras with 6mm lenses, mounted under the central microphone array provide close up views of each of the participants as shown in figure 6A. 2 Sony SSC-DC58AP CCTV cameras, each with 3.6mm semi fisheye lenses 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 6B and C).



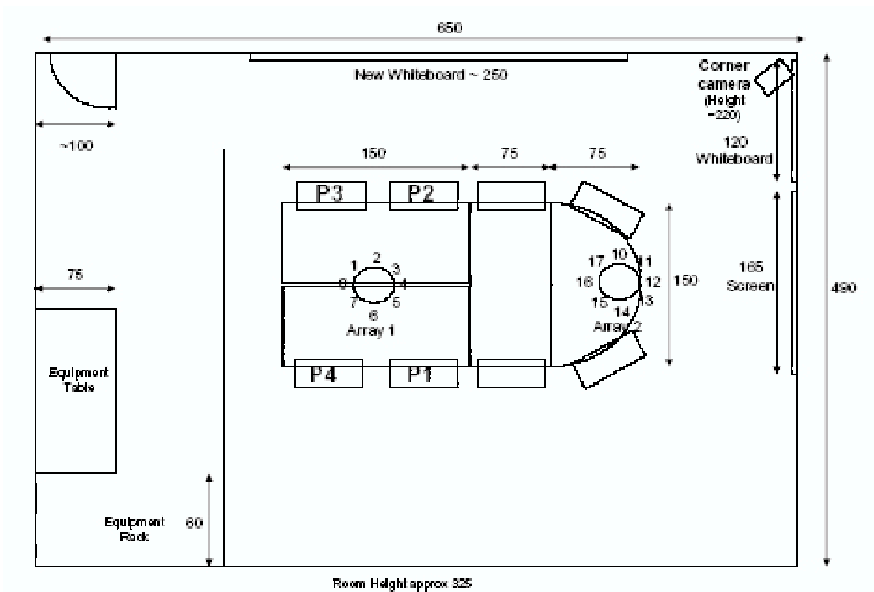


Figure 4: Overhead Schematic View of the UEDIN Instrumented Meeting Room, Showing the Participant Positions (P1-4), the Microphone Arrays, and the Wide Angle Camera Position.

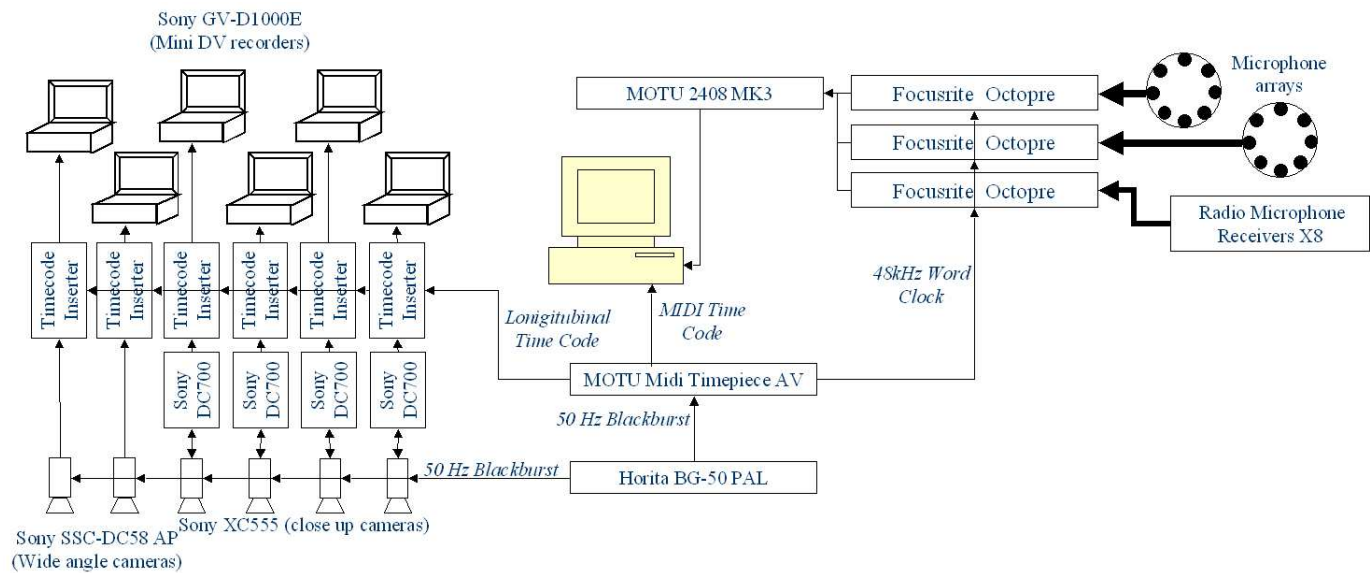


Figure 5: Diagram of the connectivity of the A/V Capture Equipment in the UEDIN room.

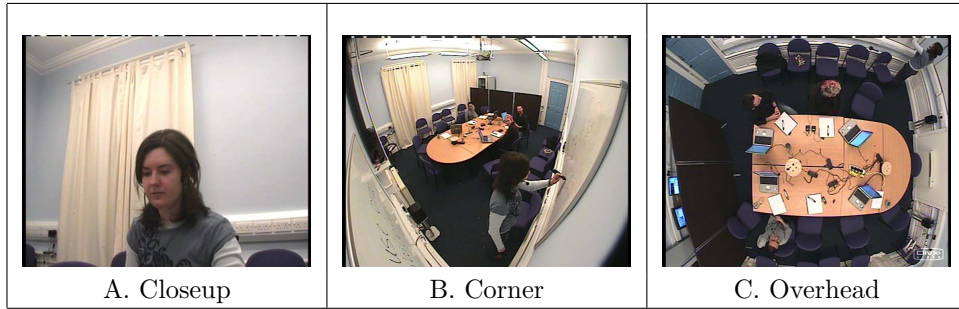


Figure 6: Camera views from the UEDIN Instrumented Meeting room.

**Mini-DV Recorders.** 6 Sony GV-D1000E digital video recorders are used to record the output of the cameras directly to Mini-DV tape. Using Mini-DV has 2 advantages - firstly it provides very reliable capture of the video with few errors or dropped frames; secondly, it provides an immediate tape back up of the raw video data. The subsequent transfer of the video to computer, as described in section 4.1, does require some manual intervention however.

### 3.1.3 Synchronization

Without further hardware to provide synchronization signals, the following errors would occur during data acquisition:

- The A to D converters in the Octopres would sample each channel at different times, resulting in time skew between audio channels.
- The cameras would acquire frames at different times, resulting in as much as 20mS difference between video channels.
- 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 used synchronization equipment as follows.

**Blackburst Generator.** The Horita BSG-50 PAL generates 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 MOTU MIDI Timepiece AV which generates all other timing signals.

**MIDI Timepiece.** The MOTU MIDI timepiece AV (MTP-AV) is capable of locking to, and generating a number of different timing signals. In the meeting room the MTP-AV locks to the Blackburst reference signal and generates :

- 48 kHz word clock. This is used to trigger the A to D converters in the 3 Octopres, ensuring that each audio channel is sampled at precisely the same instant.
- Longitudinal Time Code (LTC). This is an industry standard format where an Hours:Minutes:Seconds:Frames time code is encoded as an 80bit word for each video frame and output as a 2kHz audio signal.
- 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 Sonar recording software and used to timestamp the audio samples.

**Time Code inserters.** The Horita AVG-50 time-code inserters translate the 80bit LTC audio signal into a 90bit Vertical Interval Time Code. This 90bit code is then inserted into the top 2 lines of each video frames as a series of black and white blocks, which may subsequently be read during video playback. Since this code corresponds directly to the Midi Time Code being used to time stamp the audio recording, precise synchronization of the audio and video signals can be achieved.

#### 3.1.4 Auxiliary Data

In addition to audio and video, any auxiliary data generated by the participants whilst in the meeting is recorded. A second PC is used to capture the auxiliary data, and this uses the MIDI Time Code generated by the MTP-AV to accurately timestamp the data and ensure it is synchronized with the Audio and video streams.

**Whiteboard.** An EBEAM 2 digital white board system is used to capture any pen strokes the participants make on the whiteboard. These are stored in XML format as time stamped x-y co-ordinates of the pen.

**Beamer.** Any slides presented on the beamer are captured via a VisionRGB-Pro VGA capture card installed in the auxiliary capture PC and stored as jpeg images. Each image is time stamped using the MTC for accurate integration with other data streams.

**Handwritten notes.** Each participant has access to a Logitech I/O digital pen throughout the scenario. The pen stores the time stamped x-y co-ordinates of any pen strokes made on special paper which contains an embedded 2d bar code. The pen strokes are then downloaded to the Auxiliary Capture PC as xml files for subsequent processing. The pens are synchronised to the auxiliary capture PC at the beginning and end of each meeting, however, since they are not connected to the synchronisation equipment during the meeting, precise calibration cannot be guaranteed. In practice the pen's internal clocks do not drift by more than a few seconds during each meeting, providing sufficiently accurate calibration.

### 3.2 Deviations of other rooms

While the TNO and IDIAP rooms have essentially the same equipment, there are some minor differences in their configuration.

#### 3.2.1 The IDIAP Meeting Room

An overhead schematic diagram of the IDIAP room is shown in figure 7.

The IDIAP room has 3 wide angle video cameras rather than the two in the UEDIN room. They are positioned one at the end of the table facing the projector screen, and one behind each pair of participants, facing across the table to the opposite pair of participants.

The second IDIAP circular microphone array has only 4 elements and is mounted on the ceiling rather than on the table. In addition a binaural manikin is placed at the end of the table furthest from the screen providing 2 further audio channels.

#### 3.2.2 The TNO Meeting Room

An overhead schematic of the TNO room is shown in figure 8.

The TNO room contains a single 8 element circular microphone array mounted in the center of the table and a second, 10 element linear array mounted above the presentation screen. Each participant has a headset mounted radio microphone, but no lapel mic. The TNO room also has two wide angle cameras, one mounted above and behind the table, and one to the left hand side of the room, angled across the table. The TNO audio recording and synchronisation hardware is identical to that installed at UEDIN, however they take a different approach to the video capture. Three windows XP computers, each fitted

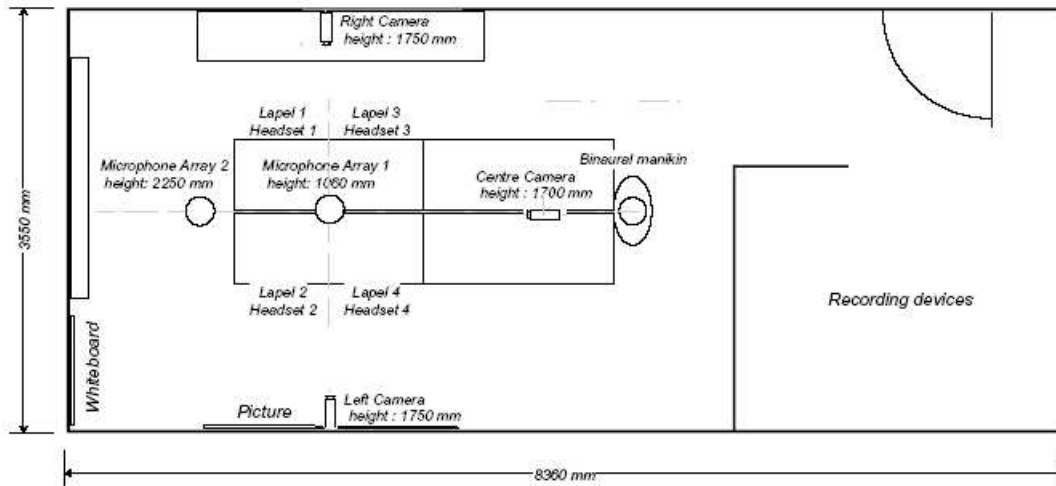


Figure 7: Overhead Schematic View of the IDIAP Instrumented Meeting Room.

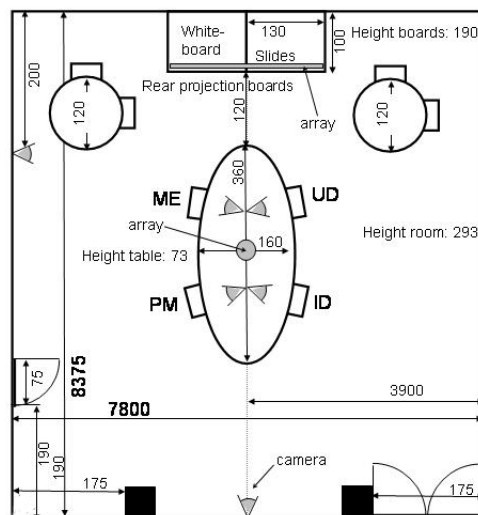


Figure 8: Overhead Schematic View of the TNO Instrumented Meeting Room.

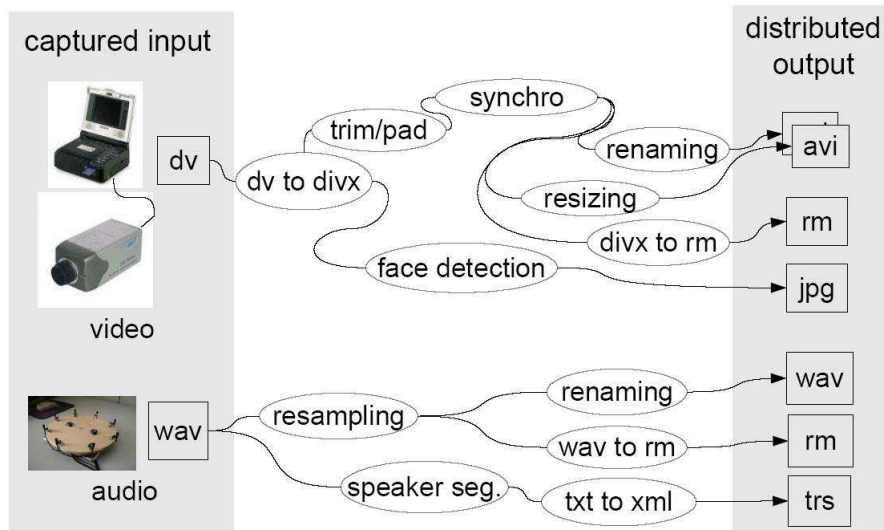


Figure 9: Audio and Video Pre-Processing

with two Osprey 210 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.

## 4 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.

### 4.1 Audio and Video Pre-Processing

Figure 9 shows the pre-processing stages for the audio and video data.

**Video** In the case of the IDIAP and UEDIN recordings, the digital video tapes are digitised via Firewire and stored to disk using the DivX AVI codec 5.2.1 The AVI's are encoded at a bitrate of 2300 Kbps with a maximum interval of 25 frames between 2 consecutive MPEG keyframes. At TNO, the encoding is part of the video capture process and so is not included as a separate stage. For synchronisation purpose, a time alignment process either trims or pads the start of the video such that the first video frame is synchronised to the first audio sample. Unfortunately, the encoding process results in some dropped frames in the video signals, which results in the video becoming desynchronised with the audio as the meeting progresses. In order to overcome this, video repair software which automatically inserts frames (chosen to be a copy of the previous frame) where frame dropping occurs is used. The video and audio therefore remain synchronised through the entire meeting.

The image resolution of these videos is high [720x576], sufficient to allow person tracking and facial feature analysis to be performed, however the file size is also relatively high. To save storage space and download times in situations where the higher resolution is not required, all the videos are also made available at a smaller size [350x280] and encoded at a lower bitrate.

Low bitrate (50Kbps) realmedia files of the combined audio and video are also produced and are available for download, or streaming from the Media File Server. All the video synchronisation software is fully documented at <http://wiki.idiap.ch/ami/SignalSynch>.

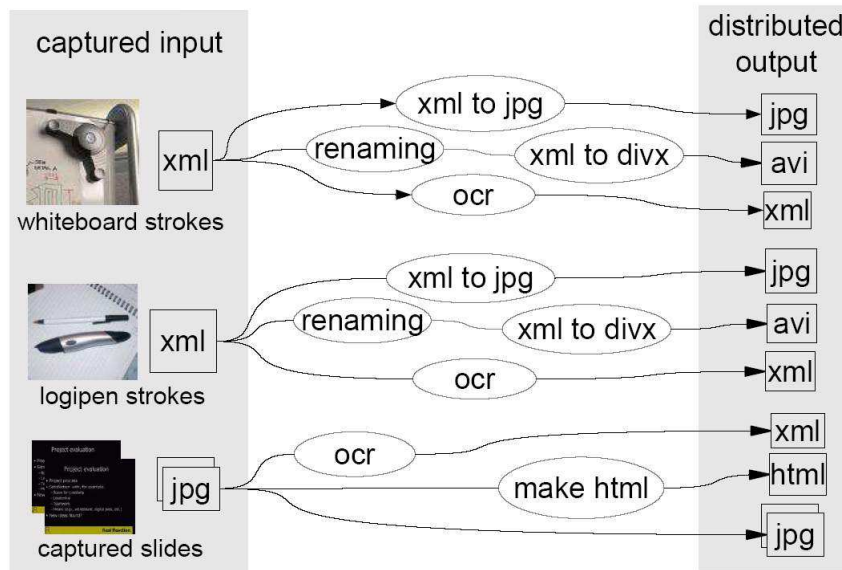


Figure 10: Auxiliary Data Pre-Processing Stages

**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 from the Media file server.

To aid in the transcription of the meetings, a simple energy-based technique was used to provide a speech/silence segmentation for each person in a meeting from their lapel microphone recordings. The technique was originally developed and tested in [5]. At each time frame, the lapel with the most energy is selected, and an automatic energy threshold is applied to classify the frame as speech or silence. The threshold is directly derived from EM training of a bi-Gaussian model on log energy values. For each person, the segmentation output (speech and silence segments) is then smoothed with a low-pass filter. The output is valid XML file in the correct format for use in channeltrans, the software used for transcribing the meetings as described in [6].

## 4.2 Auxilliary Data Pre-Processing

Figure 10 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. This shows when the strokes were made on either the whiteboard or the participants notebook. A transcription of the written data is also generated using an optical character recognition technique based on [7].

Optical character recognition based on the technique described in [8] is used to produce transcriptions of the captured slides. This, along with the captured jpeg images and html pages containing the images is made as part of the database.

## 5 Data Distribution and Management

### 5.1 The Media File server

All signal data is available on the Internet from <http://mmm.idiap.ch>. This Media File Server is the primary means for storage and distribution of multimodal meeting recordings within the project. It allows for browsing of available recorded sessions, downloading and uploading by HTTP or FTP of the data in a variety of formats, playback of media (through RTSP streaming servers and players), and it also provides web hosting and streaming servers for the Ferret meeting browser [9].

**User Interface.** The user can browse directories of meeting recordings through an initial web page that lists available meetings, has platform dependent links to the media files, shows images of participants, and indicates if speech transcripts are available. The server provides both a public space for accessing media files, and private areas available only to project members with appropriate credentials. Clicking on a particular meeting takes the user to a page with the media files for that meeting. This page contains links to all the media files (video, audio, image, XML), as well as annotations (speech transcripts, meeting agenda, speaker segmentations, etc). The meeting web interface is shown in Figure 11.

In addition to this web interface, the data files are also available through ftp, and through a cross-platform uploading facility FileManager that allows uploading of up to 200MB of media or annotation files.

**Hardware and Software.** The Media File Server hardware has been upgraded since its original implementation for the M4 (Multimodal Meeting Manager) project, to a large and stable platform [10]. Current capacity is 3 Terabytes, although this is extensible as required. This upgrade was necessary both due to the increased data rate (due to additional devices across the various AMI meeting rooms), as well as the increased load (due to the size of the consortium, and outside interest in the AMI meeting corpus).

**Mirror Site.** In addition to the principal site hosted at IDIAP, a mirror of the media file server has been established at Brno University of Technology. Having a mirror site serves two main purposes: it provides a ready off-site backup of the data, and it increases the available bandwidth for data distribution. All media and annotation files are fully mirrored, but currently the mirror only provides a simple file-list interface.



Although the Media File Server in its current version is fully operational, we note that its development is an ongoing work. Improvements are continually being implemented to respond to new media formats, facilitate integration with eventual meeting browsers, and to further improve interfaces and data management. In particular, an annotation database is being developed that will, for instance, enable search by query within annotation files to improve access to the underlying media files.

### 5.2 The AMI Dataflow Wiki

Due to the complexity and distributed nature of the data collection, a section of the AMI wiki (<http://wiki.idiap.ch/ami/AmiDataFlow>) has been used to co-ordinate and exchange information concerning the collection and post processing activities. The Wiki contains information of value to both data providers and consumers such as : descriptions of meetings and their IDs; Issues concerning data collection for specific meetings, such as audio drop-outs or loss of synchronisation; data pre-processing issues; definitions of data sub-sets, such as an initial hub-set used to trial annotation schemes; descriptions of non-scenario data recordings to give them some context. It is intended that information from the Wiki will become the primary reference for data consumers and will provide full details of the finalised database.

HELP | MMM Home

**AMI scenario Hub -- meeting IS1008c** **10-12-2004** **13h59**




Press  to start Ferret Meeting Browser.

**Side cameras** ⊙

R

L

C


**Close Ups** ⊙

Closeup4

Closeup2

Closeup1

Closeup3



**Audio**

<input type="button" value="Play"/> Array1-1 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Array1-2 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Array1-3 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Array1-4 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Array1-5 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Array1-6 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Array1-7 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Array1-8 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Array2-1 <a href="#">.wav</a> <a href="#">.m</a>
<input type="button" value="Play"/> Array2-4 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Headset-1 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Headset-2 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Headset-3 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Headset-4 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Lapel-1 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Lapel-2 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Lapel-3 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Lapel-4 <a href="#">.wav</a> <a href="#">.m</a>
<input type="button" value="Play"/> Array2-2 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Array2-3 <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Manikin-Left <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Manikin-Right <a href="#">.wav</a> <a href="#">.m</a>	<input type="button" value="Play"/> Mix-Headset <a href="#">.wav</a> <a href="#">.m</a>	<input checked="" type="button" value="Play"/> Mix-Lapel.wav <a href="#">.wav</a> <a href="#">.m</a>			

**Annotations**

- [IS1008c-speakersegAll.xml](#)
- [IS1008c-speakersegChan0.xml](#)
- [IS1008c-speakersegChan1.xml](#)
- [IS1008c-speakersegChan2.xml](#)
- [IS1008c-speakersegChan3.xml](#)

**Transcripts**

- [IS1008c-transcript-new.html](#)

**Other Files**

*Documents:*

- [IS1008c-Transcript.trs](#)

*Handwriting data:*

- [IS1008c-Pen\\_3\\_Page070-10.12.2004.pen](#)
- [IS1008c-Pen\\_4\\_Page073-10.12.2004.pen](#)
- [IS1008c-Pen\\_1\\_Page067-10.12.2004.pen](#)
- [IS1008c-Pen\\_2\\_Page072-10.12.2004.pen](#)

*Slides:*

- [slide navigation](#)

*Whiteboard strokes:*

- [IS1008c-strokes.xml](#)

To download files, *right-click-save* on the blue links, or FTP from <ftp://mmm.idiap.ch/private/amiHub/157119212>.

Figure 11: Available meeting data files as displayed on MMM.



## 6 Current Status of Hub Corpus

A total of approximately 90 hours of data has been recorded to date. This data consist of 71 hours of design scenario data, 5.5 hours of less controlled elicitations (as described in 1) and 13 hours of 'real' naturally occurring meetings.

**Design scenario meetings.** Of the 71 hours of scenario meetings, 30 hours have been recorded in the UEDIN room (meetng IDs ES20\*), 20 hours have been recorded in the IDIAP room (meeting IDs IS10\*) and 21 hours have been recorded in the TNO room (meeting IDs TS30\*)

**Other elicited meetings** These meetings constitute 5.5 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 13 hours of recordings made in the UEDIN room (meeting IDs EN\*). 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.

At time of writing, all the audio and video data for the UEDIN and IDIAP Scenario meetings is available for download and streaming from the MMM server (<http://mmm.idiap.ch/protected>), as are the IDIAP IB\* series of non scenario meetings. The TNO scenario meetings and UEDIN non scenario meetings are currently undergoing the Pre-processing described in section 4.1, however we foresee no problems with this process and anticipate the remainder of the recorded data to be made available shortly.

We plan to record further real meetings and are actively seeking groups who hold regular meetings within within the organisations equipped with instrumented meeting rooms. These will be made available as and when they are held, and as such the data collection is an ongoing effort.

## 7 Spoke data recording in BUT mobile meeting room

In addition to the core data recorded in AMI, "spoke" data is being recorded in specific meeting room set-ups. One of them is the mobile meeting room at Brno University of Technology (BUT) depicted in Fig. 12. Its setup is low-cost, mobile, built from off-the-shelve products, and easy to install and operate. The specialty of our meeting room is the 360 degree image capturing using hyperbolic mirror, with the following image-processing.



Figure 12: Mobile meeting room.

## 7.1 Hardware

The *audio* is captured by 4 lapel microphones Sennheiser MKE 2 P-C (same as in IDIAP). 2 mixers Behringer EURORACK UB1204-PRO are providing gain control and phantom power supply (actually 90% of mixers are not used, but they are cheaper than special mike pre-amps with phantom feeders). The notebook PC is equipped by two Hi-Fi PC-MCIA sound-cards VX-pocket. *Video* is taken by digital camera SONY CLIP 345. The hyperbolic mirror by NEOVISION (Prague) is used to transform the 360 degree view for the camera. The PC is connected through IEEE1394 (FireWire) interface. The camera also provides 2 additional (global) audio channels. The total *cost* of the hardware (except the notebook) is  $\approx 4.2$  kEUR. The hardware is easily transportable by one person the only bulky thing being the stand for the camera.

Recently, the hardware was modified to further compact the meeting room and to obtain better resolution for video processing. The mixers and sound-cards are replaced by a FireWire sound card RME FireFace 800. The original camera is replaced by a HDTV camera SONY HDR-FX1 (Fig. 13, left) with  $1440 \times 1080$  pixel resolution.



Figure 13: HDTV camera and the original image.

## 7.2 Recording software

The recording is performed using the DV-Capture tool developed at BUT. The tool is based on Direct-show libraries and runs under MS-Windows. All audio and video streams are stored in real time: two audio wav files with four audio channels and one DV-compressed video avi file (Fig. 13, right) with video and additional stereo audio stream from camera. The audio and video channels are acquired synchronously using the Direct-X interface. To ensure minimum time-shift between the channels, low-level layers were used and on top of this, time differences of channels are monitored to allow possible further resampling of the output channels.

## 7.3 Post-processing

The camera mounted to the holder with mirror is susceptible to vibrations. These vibrations create undesirable movements in the transformed image, which is increasing in the upper border direction. Automatic detection of the image parameters can solve this problem. The output image from catadioptric

system has circular shape, which is given by the mirror top view. The transformation algorithm uses information about circle center and radius. Because input image consists of simple background and circular “omni image”, it is easy to find border of the “omni image” with its properties. In this way, it is possible to stabilize the image with sub-pixel of  $\pm 0.2$  pixels [11], the method runs also in real-time.

Further post-processing is needed to unwrap video to panoramic format of view (Fig. 14) [12].

The audio is pre-processed by speaker-turn segmentation. As the only microphones in this meeting room are lapel ones heavily suffering from cross-talk, we use a technique which takes into account cross-correlations, values of its maxima, and energy differences as features to identify and segment speaker turns [13].



Figure 14: Transformed image.

## 7.4 Status

7 short “real” meetings with the total length of about 2 hours were recorded in the HDTV-version of the meeting room. These data were mainly used for work on video tracking and automatic video editing algorithms [11, 14]. Recording of 20 hours of data is planned.

## References

- [1] McGrath, J.E., Hollingshead, A.: *Interacting with Technology: Ideas, Evidence, Issues and an Agenda*. Sage Publications, Thousand Oaks (1994)
- [2] 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
- [3] 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)
- [4] Pahl, G., Beitz, W.: *Engineering design: a systematic approach*. Springer, London (1996)
- [5] Lathoud, G., McCowan, I.A., Odobez, J.M.: *Unsupervised location-based segmentation of multi-party speech*. In: *ICASSP-NIST Meeting Recognition Workshop*, Montreal (2004) <http://www.idiap.ch/publications/lathoud04a.bib>.
- [6] ICSI: *Extensions to transcriber for meeting recorder transcription*. <http://www.icsi.berkeley.edu/Speech/mr/channeltrans.html> (2003)
- [7] Liwicki, M., Bunke, H.: *Handwriting recognition of whiteboard notes*. In Marcelli, A., ed.: *12th Conference of the International Graphonomics Society*, Salerno (2005)
- [8] Chen, D., Odobez, J.M., Boulard, H.: *Text detection and recognition in images and videos*. *Pattern Recognition* (2004)

- [9] Wellner, P., Flynn, M., Guillemot, M.: Browsing recorded meetings with Ferret. In Bengio, S., Bourlard, H., eds.: Machine Learning for Multimodal Interaction: First International Workshop, MLMI 2004, Martigny, Switzerland, June 21-23, 2004, Revised Selected Papers. Lecture Notes in Computer Science 3361. Springer-Verlag, Berlin (2005) 12–21
- [10] Formaz, F., Crettol, N.: The IDIAP multimedia file server. Technical Report IDIAP-Com 04-05, IDIAP (2004)
- [11] Sumec, S., Potucek, I., Zemcik, P.: Automatic mobile meeting room. In: Proceedings of 3IA'2005 International Conference in Computer Graphics and Artificial Intelligence. (2005)
- [12] Nayar, S., Baker, S.: A theory of catadioptric image formation. Technical Report CUCS-015-97, Department of Computer Science, Columbia, University, (1997)
- [13] Motlicek, P., Burget, L., Cernocky, J.: Non-parametric speaker turn segmentation of meeting data. In: Eurospeech 2005, Lisbon. (2005)
- [14] Sumec, S.: Multi-camera automatic video editing. In: CCVG 2004, Warsaw, Poland. (2004)