AMI CONSORTIUM

# AMI & AMIDA Projects

## 2004-2010

www.amiproject.org

# Final Newsletter

# Final Report

01.01.2010

Sixth Framework Programme

# Last AMIDA Newsletter (4 February 2010)
# Preface

Over the last six years, two European Integrated Projects, referred to as AMI (Augmented Multiparty Interaction, January 2004-December 2006, EC funding=8,800KEuros, total budget=16'800KEuros) and its follow-up project AMIDA (Augmented Multiparty Interaction with Distance Access, October 2006-December 2009, EC funding=9,900KEuros, total budget=13,500KEuros), have set up serious grounds in multiple research areas related to human-human interaction modeling, computer enhanced human-human communication (especially in the context of face-to-face and remote meetings), social communication sensing, and social signal processing.

As described in more detail in the AMI/DA web site (http://www.amiproject.org), the development of such technologies, their related applications, and user-centric evaluations, require to go beyond, and integrate in a principled way, the state-of-the-art in several multi-disciplinary areas, including models of group dynamics (behavioral and social sciences), audio and visual processing and recognition, models to combine multiple modalities, the abstraction of content from multiparty meetings, and issues relating to human-computer interaction. These R&D themes are underpinned by the ongoing capture of user requirements, the development of a common infrastructure, and evaluations of the resultant systems. While meetings provide a rich case study for research, and a viable application market, many of the scientific advances that have been made within the AMI and AMIDA projects are wider than any single application domain. Each of the technologies briefly discussed in this report has broad application potential, for example in security, surveillance, home care monitoring, and in more natural human-computer interfaces. Progress on several fronts were necessary to develop the targeted meeting support technologies, from establishing the necessary framework for successful long-term collaborative research, through to development of multiple prototypes to access (online or offline) meeting archives (including meeting browsers and automatic content linking systems). A common hardware and software infrastructure had to be established, and within this we undertook ambitious data collection and annotation efforts. We have also developed leading edge technologies in audio, visual and multimodal processing, and in content abstraction.

Besides numerous scientific reports and large amounts of high quality publications, the AMI and AMIDA projects also paid particular attention to knowledge transfer, marketing and information, as well as in new approaches towards technology transfer. Our main communication tool was our quarterly newsletters, all available from our web site, and which were being sent by mail or email, respectively to about 100 and 300 people. The current issue is thus the last one of a long series. While focusing again on some recent technology transfer successes resulting of the projects, it is also complemented by a general overview of the six years of work in the targeted area. While being easily accessible to non-specialist, this overview will also be one of the (introductory) chapters of an AMI/DA book currently being finalized.

Finally, the AMI/DA projects have also dedicated attention and resources to the appropriate transfer of technology which can be identified on the basis of research conducted by the project consortium members. Our activities in this area included communication internally with scientists, research managers and technology transfer offices, and externally, with commercial companies and a variety of organizations and individuals who influence the direction of future commercial products for business meetings. The efforts were mainly concretized by the set up of a large AMI/DA Community of Interest, which members were involved in several mini-projects, as also described in the business portal of the web site.

We thank you all for having followed our progress over the last several years, often with great interest, to have collaborated with us, and we hope that the end of this project and these newsletters will not mean the end of our collaboration.

AMI CONSORTIUM

Issue N° 20 / February 2010

Editor : Valérie Devanthéry, valerie.devanthery@idiap.ch

# Newsletter

www.amiproject.org

AMI c/o Idiap Research Institute, Centre du Parc,
Rue Marconi 19, P.O. Box 592, CH-1920 Martigny
info@amiproject.org - www.amiproject.org

## Contents

## News

### ICMI-MLMI 2010

Nov 8-12, 2010
Beijing China

The Twelfth International Conference on Multimodal Interfaces and the Seventh Workshop on Machine Learning for Multimodal Interaction

*http://www.acm.org/icmi/2010*

### Final Review Meeting

February 18-19, 2010
Edinburgh, United Kingdom

The meeting to prepare the final review will be held in Edinburgh on February 17th 2010, one day before the the final review meeting.

© AMI Consortium, all rights reserved

Cover Story

## Activity Recognition for Eating Scenarios

THE TECHNISCHE UNIVERSITÄT MÜNCHEN AND NOLDUS INFORMATION TECHNOLOGY

### Overview

For many applications such as customer research or service improvement in restaurants it would be beneficial to observe the people in a restaurant and know exactly when they eat, drink, what they have on their table, etc. Noldus Information Technology is working in this field and has with the Observer XT already a successful product on the market.

However, until now videos still have to be manually labeled. Of course, the industry has a huge desire to let a computer perform this task automatically. However, the exact automatic recognition of the eating activities is challenging.

The Technische Universität München (TUM) had a coorperation with Noldus Information Technology on the subject. Eleven sample videos showing videos of persons eating and drinking were analyzed.

Low level information was extracted by seven modules namely face detection/tracking module, hand detection/tracking module, table detection module and object detection/tracking module, where an object can be a glass or a plate. All modules are implemented based on state-of-the-art algorithms.

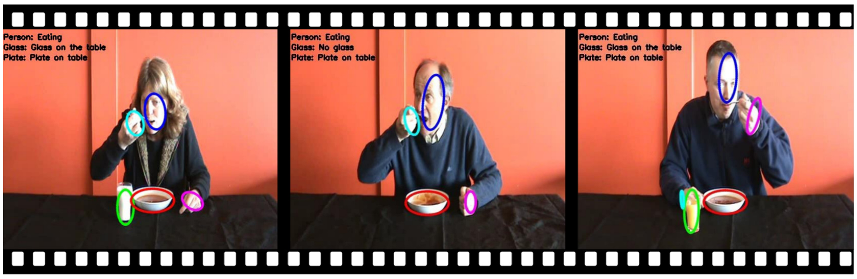
### Activity Recognition

In order to perform activity recognition the information obtained by the recognition modules are gathered and evaluated. Therefore, we determined three objects of interest: the person, the glass, and the plate. All of those objects of interest can be modeled as a finite state machine.

The person can be in one of the following states: No person, coming in, sitting, drinking, eating, or going out. A glass or a plate can be in the states: no object, on table, or moving. Below some screenshots from our activity recognition program are shown. The program generates an output file that can be directly imported into the Observer XT Software. All eleven video sequences have been manually labeled and this ground truth has been used to obtain quantitative results for our algorithm, which are shown in the table below.

| [%] | Glass | Plate | Person |
|---|---|---|---|
| Correct Frames | 83.00 | 91.80 | 83.16 |
| Wrong Frames | 17.00 | 8.20 | 16.84 |
| Best | 98.41 | 97.82 | 91.20 |
| Worst | 59.34 | 63.97 | 74.63 |

Dipl.-Ing. Moritz Kaiser, TUM
*moritz.academic@googlemail.com*



Augmented Multiparty Interaction Distance Access (AMIDA) is an Integrated Project funded by the EC's 6th Framework Program, jointly managed by Idiap (CH) and the University of Edinburgh (UK).

## Engagement and Floor Control in Hybrid Meetings

BY THE HUMAN MEDIA INTERACTION GROUP OF THE UNIVERSITY OF TWENTE

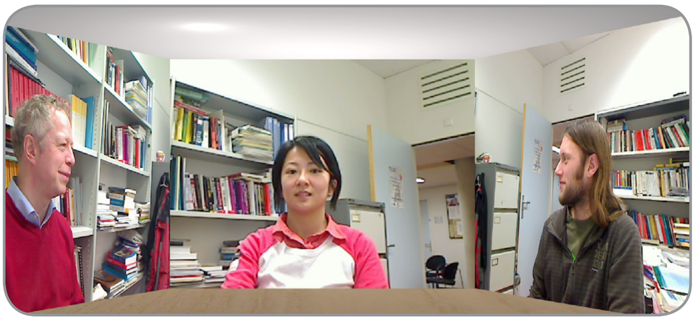


*Figure: showing the 3D interface for the remote participant*

It is a frequently observed fact that remote participants in hybrid meetings often have problems to follow what is going on in the (physical) meeting room they are connected with. From extensive analyses of face to face and remote meetings it is clear how important non-verbal social behavior is in communicating who is being addressed and who is expected to take turn. One of the possible applications of AMIDA research in real-time automatic scene analyses and meeting behavior recognition is in the development of technology and interfaces that support participants in distributed meeting environment.

The Human Media Interaction group of the University of Twente has developed a User Engagment and Floor Control Demonstrator, a system that uses modules for online speech recognition, real-time visual focus of attention as well as a module that signals who is being addressed by the speaker. A built-in keyword spotter allows an automatic meeting assistant to call the remote participant's attention when a topic of interest is raised, pointing at the transcription of the fragment to help him catch-up. The first version of the UEFC demo was presented at the AMIDA review meeting in Edinburg last year.

In the final year of AMIDA we focussed on two main tasks. The first task is the integration of the UEFC demo and the Content Linking Demo. This has resulted in an offline version of the UEFC demo that demonstrates how both demonstrators can work with the same instance of the HUB database that contains the annotation layers of the meeting as well as the documents that are automatically retrieved based on a set of key phrases selected by the user.

Both the offline and the online version of the UEFC demo make use of the HMI Media Streaming package that handles the synchronisation, compression and streaming of video and audio for communication, processing and recording. The HMI Media Streaming software has also been used in the remote meeting experiments that were performed in an AMIDA Miniproject with TXchange, a member of the COI (see a previous issue of this Newsletter). The HMI media streaming package, based on DirectShow for MS Windows, allows easy development of user interfaces in a modular way.

The second and main task that has been performed is to experiment with two different user interface for videoconferencing: a classical 2D interface and a 3D interface. The experiments were performed in the meeting room of the new SmartXP Lab at the University of Twente. The main differences between both experimental conditions is that in the 3D version non-verbal communicative behavior in the form of gaze is transmitted in a substantially improved way. We compare a conventional video conferencing interface versus an interface where video streams were presented to remote participants in an integrated 3D environment, in a context where a small group of three co-located persons had a meeting joined by one person located on a remote place. The conventional interface we used was in essence like a Skype or Adobe connect interface in the sense that meeting participants are visible in separate video frames, and each of them would be looking straight into their own webcam.

In both conditions the video image of the remote participant was visible to the co-located persons on a classical, but large, video screen. Such multimodal interfaces offer already a lot in that both speech as well as facial expressions are communicated. But certain non-verbal behavioral aspects are still lacking, in particular body pose and gaze direction from one person to another. Our integrated interface tries to improve these aspects, aiming at enhancing the presence of participants. On the remote participant side it employs a basic 3D interface where the (video images of) other participants appear to be sitting around a virtual table, consistent with the real, physical situation (see Figure: showing the 3D interface for the remote participant). Cameras capturing each of the co-located participants were no longer in front of them, but rather repositioned, so that looking in the direction of the remote participant (screen) would coincide with looking into the camera, as far as possible.

The main intended effect of the 3D interface is that it becomes much easier to observe who is looking to whom. We arranged an experiment where ten groups, each consisting of four participants, held short meetings using the two different interfaces, while being observed by means of cameras as well as through a two-way mirror, allowing observers to track the participants gaze behavior, in order to measure the amount of attention that participants receive from others, and to analyze turn taking behavior. We also measured perceived presence by means of a social presence questionnaire, asking about aspects like (perceived) co-presence, message understanding, and attention allocation. Finally, we held interviews immediately after meeting sessions. The first results show interesting and statistically significant differences between the two experimental situations, mostly favoring our integrated interface. Ambiguity with respect to who is looking to whom in the classical interface has been observed, and might at least partially explain these results. We expect that the corpus of synchronized audio and video recordings, together with the questionnaires are a valuable data set for future research in remote meeting behavior.

*Rieks op den Akker, Job Zwiers, Hendri Hondorp, Betsy van Dijk, Olga Kulyk, Dennis Hofs, Anton Nijholt, Dennis Reidsma,*
*Human Media Interaction, University of Twente,*
*Enschede the Netherlands*
*infrieks@cs.utwente.nl*

AMI CONSORTIUM

Issue N° 20 / February 2010

Editor : Valérie Devanthéry, valerie.devanthery@idiap.ch

Newsletter

# Meeting Profiler and Expert Finder: two applications from WP5

ERIK BOERTJES, WESSEL KRAAIJ, STEPHAN RAAIJMAKERS, CORNÉ VERSLOOT, JOOST DE WIT

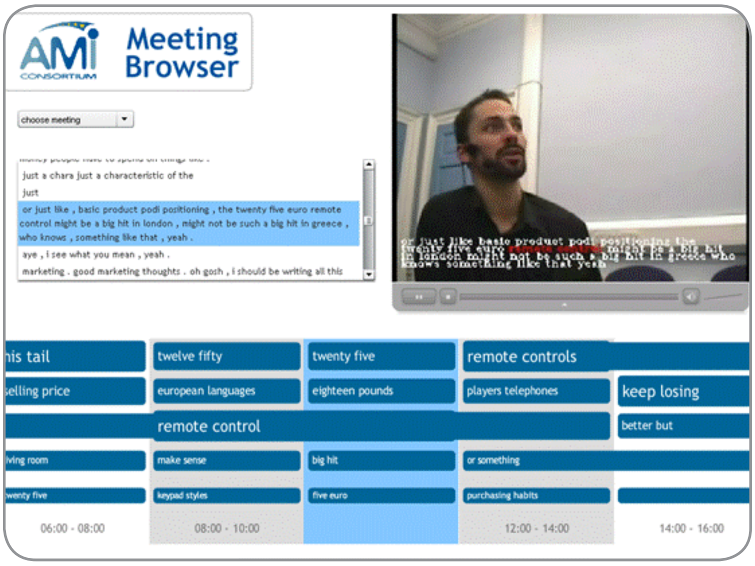
## Meeting Profiler

The meeting profiler enables the user to quickly view what topics were addressed during a meeting and at what time. Its interface is shown in Figure 1. On the right hand side is the video window, containing the footage of the selected meeting. The transcript in the transcript window (left of the video) is aligned with the chosen point in time in the video. The Tagcloud window (below) always highlights the tag cloud corresponding with the point in time in the video allows for quick browsing through the meeting.

The Meeting profiler makes use of the Tag Cloud generator component from WP5. This component automatically generates tag clouds (sets of weighed terms) by:

1. Splitting the meeting transcript into equal pieces (windows) of a predefined length.
2. Extracting the most salient terms from each window using language models.
3. Selecting the top N terms to be placed on the timeline.

Another key component from WP5 used in the Meeting Profiler is the Video Editor which automatically produces one video from the various sources. The video that is produced for the visualization of a meeting takes as input all available footage, consisting of close-up videos of every participant, and an overall global view video of the meeting. A specially designed automatic video producer makes an intelligent decision as to which video stream to mix in the final mix, depending on who is speaking. The final video is synchronized with the text transcript of the meeting.



just a chara just a characteristic of the

just

or just like , basic product podi positioning , the twenty five euro remote control might be a big hit in london , might not be such a big hit in greece , who knows , something like that , yeah .

aye , i see what you mean , yeah .

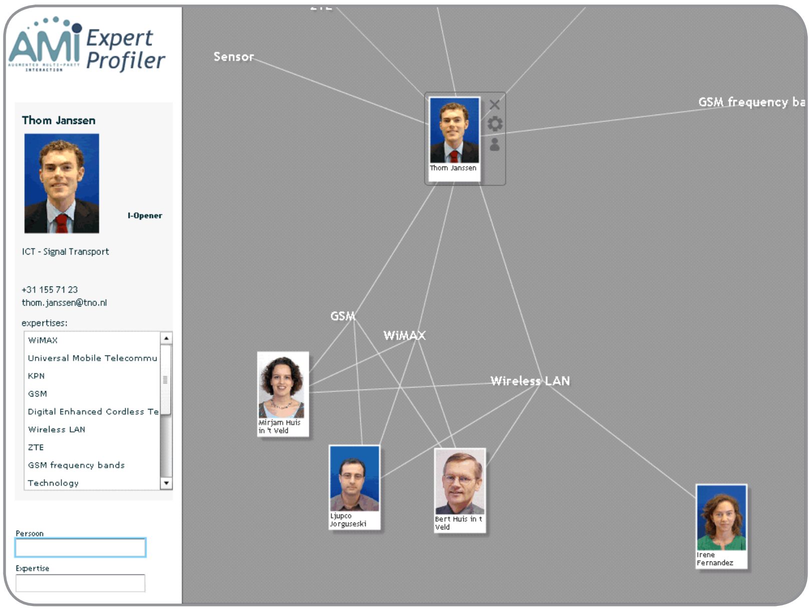
marketing . good marketing thoughts . oh gosh , i should be writing all this

or just like basic product podi positioning the twenty five euro remote control might be a big hit in london might not be such a big hit in greece who knows something like that yeah

Figure 1: Meeting Profiler interface

## Expert Finder

Finding the person with the right expertise for a project can be difficult in large organizations, especially if they consist of several geographically distributed offices. Existing expertise-finding methods often do not sufficiently succeed in solving this problem. We therefore implemented the Expert Finder (see Figure 2), a system that automatically builds expert profiles of people and provides for more flexible ways of searching than traditional tools.



Thom Janssen

I-Opener

ICT - Signal Transport

+31 155 71 23
thom.janssen@tno.nl

expertises:
WiMAX
Universal Mobile Telecommu
KPN
GSM
Digital Enhanced Cordless Te
Wireless LAN
ZTE
GSM frequency bands
Technology

Persoon

Expertise

Figure 2: Expert Finder interface

The Expert Profiling component from WP5 forms the bases for the Expert Finder. It takes the name of a person and automatically retrieves documents of which the person is the author, the co-author or is associated with in another way, as expressed by the metadata of the document. Documents are retrieved from both public sources on the internet (like Google Scolar, Altavista, etc.) and local sources on a company intranet. They are analyzed by a topic spotter that derives terms that describe expertise.

Instead of the query-result paradigm, we chose to offer the user the possibility of browsing through a network of persons and their expertise allowing for so called 'exploratory search', in which a user intuitively jumps from one node to the other, expanding some nodes and deleting others until the suitable persons have been found.

*Erik Boertjes, Wessel Kraaij, Stephan Raaijmakers,*
*Corné Versloot, Joost de Wit*
*TNO*
*erik.boertjes@tno.nl*

AMI CONSORTIUM

Issue N° 20 / February 2010

Editor : Valérie Devanthéry, valerie.devanthery@idiap.ch

Newsletter

# The Automatic Content Linking Device (ACLD): a project-wide demonstrator

## A JOINT ACHIEVEMENT OF THE AMIDA PARTICIPANTS

The ACLD is a meeting assistant that provides just-in-time access to potentially relevant documents or fragments of past recorded meetings, based on speech from an ongoing discussion. Participants in meetings often mention such documents, but do not usually have the time to search for them. Therefore, the ACLD monitors their verbal output and retrieves documents that might be relevant to them, from several repositories containing past meeting recordings, related documents, slides, minutes, etc., or from the Web. The ACLD is intended to be used during meetings, but the system can also be used on a past meeting (e.g. from the AMI Meeting Corpus) or a presentation (e.g. a recorded talk or course) by replaying the corresponding recording, and pointing the ACLD to a repository containing related media and documents. The ACLD is the first such system that is fully implemented in a multimodal interaction and archival context.

The architecture of the ACLD includes the following main modules, which can be activated according to various scenarios through a System Controller. The Document Bank Creator and Indexer extracts text from documents and builds an index for subsequent retrieval. The Query Aggregator submits queries based on spoken input at regular intervals during a meeting and aggregates the results.

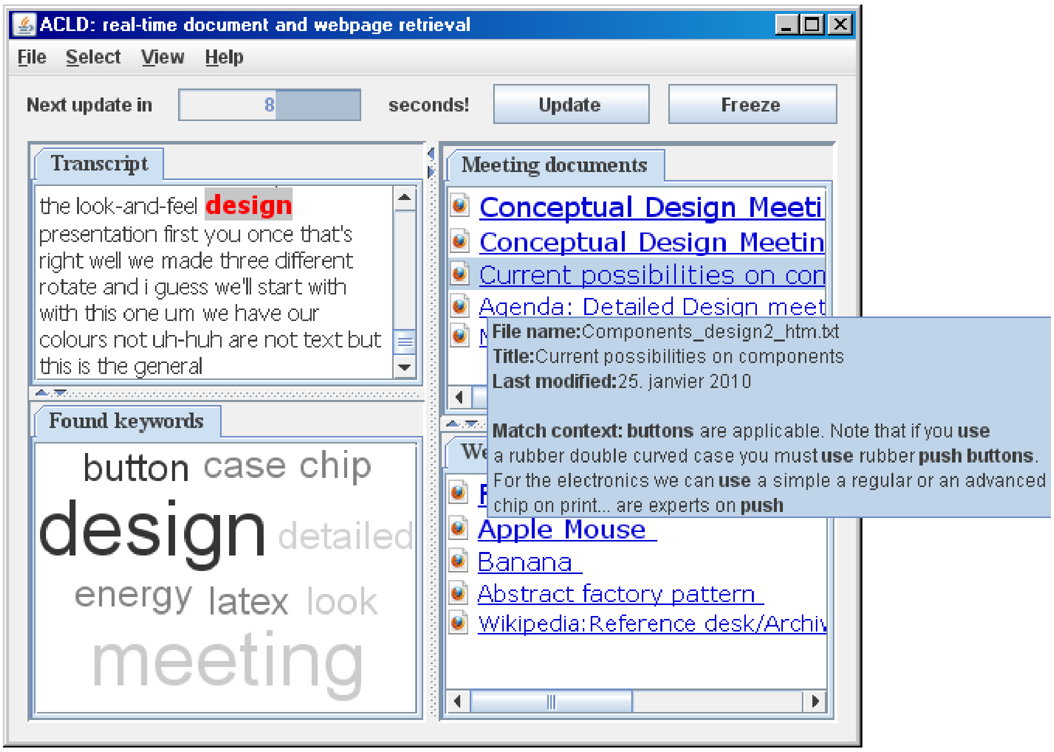
The User Interface displays the current search results as clickable document links, and provides through these links access to past documents, meeting recordings, and web pages. Input processing modules from the AMI Consortium can be used to process the incoming verbal input: automatic speech recognition, keyword spotting, disfluency removal, and dialogue act segmentation. The modules exchange data using the Hub, a client-server architecture developed within AMIDA, and audio and video signals are broadcast thanks to the HMI Media Server used in the User Engagement and Floor Control device.

The latest improvements of the ACLD have focused on a new, modular user interface, a more powerful system controller, and the extension of the document repositories and formats to which the ACLD gives access. Moreover, integration with the processing modules and portability were improved, and feedback from potential users was collected in several settings.

The new Modular UI contains five widgets or tabs (four of which are shown in the Figure 1 ), which can be enabled, disabled, or arranged at will.

Counterclockwise from upper left, these widgets show:

(1) all the words recognized by ASR with highlighted keywords;
(2) a tag-cloud of the detected keywords, coding for recency and overall frequency of mention;
(3) labels of relevant web links found (via Google) within a web domain that can be specified from the menu, with relative relevance indicated by position, font size, and emphasis; and
(4) labels of the relevant documents and/or past meeting snippets found in the meeting index, with similar coding of their relevance.



the look-and-feel design presentation first you once that's right well we made three different rotate and i guess we'll start with with this one um we have our colours not uh-huh are not text but this is the general

Conceptual Design Meeti
Conceptual Design Meetin
Current possibilities on con
Agenda: Detailed Design meet

**File name:**Components_design2_htm.txt
**Title:**Current possibilities on components
**Last modified:**25. janvier 2010

**Match context: buttons** are applicable. Note that if you **use** a rubber double curved case you must **use** rubber **push buttons**. For the electronics we can **use** a simple a regular or an advanced chip on print... are experts on **push**

Apple Mouse
Banana
Abstract factory pattern
Wikipedia:Reference desk/Archiv

*Figure 1: The ACLD user interface*

A widget showing results from Google Desktop is also available. As shown in the Figure, hovering over the search results displays relevant excerpts from the documents, while clicking on a name opens the corresponding document in an appropriate editor or meeting browser.

The ACLD is a joint achievement of the AMIDA participants. Design and development were mainly done at Idiap, the U. of Edinburgh, and DFKI, with contributions from TNO, the U. of Twente, the T.U. of Brno, and other partners.

*Andrei Popescu-Belis*
*Idiap Research Institute*
*andrei.popescu-belis@idiap.ch*

AMI CONSORTIUM
Issue N° 20 / February 2010
Editor : Valérie Devanthéry, valerie.devanthery@idiap.ch
Newsletter

## Integrating AMI ASR technology into a lecture retrieval system

### MINI-PROJECT BETWEEM KLEWEL AND IDIAP RESEARCH INSTITUTE

Although the AMI and AMIDA projects have focussed on meeting analysis, the technology is by no means confined to meetings. One example is the automatic speech recognition (ASR) component. Generally, the ASR is an adaptive system, trained for open-vocabulary and non-native English speakers. The hypothesis was that lectures also fitted this general description. In many ways, the lecture environment is easier: One person speaking clearly with little background interference. In another sense, lectures pose their own difficulties: Distant microphones, reverberation and potentially quite task specific vocabulary.

Klewel is well known to the AMI community. Over the past few years, Klewel has recorded data from a variety of different lecture scenarios. Nowadays, Klewel is able to offer such lectures in a browsable format, and is becoming recognised as a world-leader in such lecture recording and archiving. There is a strong desire from both Klewel customers and technologists for ASR in the Klewel system. Such ASR would allow captioning and searching.

To evaluate the AMI system in the lecture scenario, approximately one hour of lecture material provided by Klewel was annotated. The lectures were also recognised by various incarnations of the AMI ASR system. The recognition results could then be compared to the annotation ground truth. The results were extremely encouraging. The word error rates vary: Using simple features and a real-time decoder yields around 50% error rate. More sophisticated features in a multi-pass confuguration give closer to 25% error rate. To put this in perspective, these figures agree closely with those that we obtain on meetings.

To evaluate how well these numbers translate into search performance, spoken term detection (STD) experiments were also conducted. Results were equally encouraging: The equal error rate false alarm / miss rate was 7.4%. Put another way, this means that searches are more than 90% successful.

In parallel to the numerical investigations at Idiap, Klewel developed a demonstration interface for their meeting browser. The browser can read ASR results in an agreed file format, and do two distinct things:

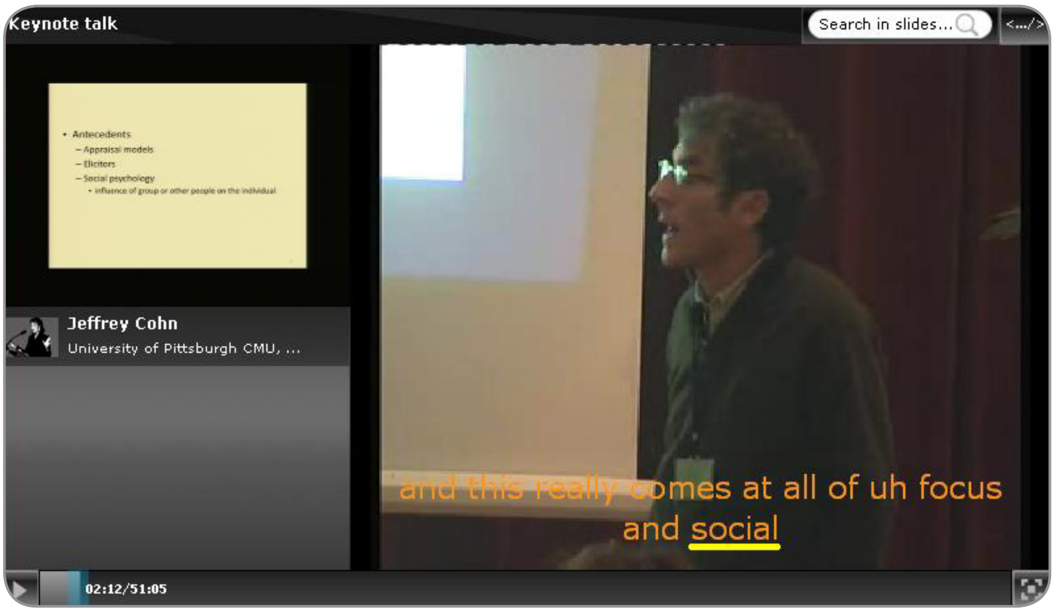
1. It can render the results as closed captions.
2. It can search the lecture based on keywords.

The description is brief, this is because the browser is able to speak for itself: Just go to http://www.klewel.com/amida_asr_demo/ and experience it for yourself.

The mini-project raised, but was only able to partially address a complicated IP situation. The AMI ASR group are now following this up with work on the IP side of commercialisation. This presents a difficulty, in that the IP is distributed both within and outside AMI partners, but also an opportunity for us to consolidate all this in specialised entities. Watch this space!

A technical report based on the AMIDA/Klewel experiments is openly available here: http://publications.idiap.ch/downloads/reports/2010/Motlicek_Idiap-RR-03-2010.pdf

The project was undertaken by Petr Motlicek and Phil Garner from Idiap and Mael Guillemot and Vincent Bozzo from Klewel.



and this really comes at all of uh focus and social

*Klewel demonstration interface*

*Phil Garner*
*Idiap Research Institute*
*Phil.Garner@idiap.ch*

Issue N° 20 / February 2010

Editor : Valérie Devanthéry, valerie.devanthery@idiap.ch

AMI CONSORTIUM Newsletter

# Events

## ICMI-MLMI 2010

### Nov 8-12, 2010, Beijing China

The Twelfth International Conference on Multimodal Interfaces and the Seventh Workshop on Machine Learning for Multimodal Interaction will be held jointly in Beijing China during November 8-12, 2010.

The main aim of ICMI-MLMI 2010 is to further scientific research within the broad field of multimodal interaction, methods, and systems, focusing on major trends and challenges, and working towards identifying a roadmap for future research and commercial success.

**Important Dates**

- April 1, 2010: Workshop proposals
- May 1, 2010: Workshop proposal acceptance notification
- May 20, 2010: Paper submission
- July 20, 2010: Author notification
- August 20, 2010: Camera-ready due
- Nov. 8-10, 2010: Conference
- Nov. 11-12, 2010: Workshops

More information:
*http://www.acm.org/icmi/2010*

*Valérie Devanthéry*
*Idiap Research Institute*
*valerie.devanthery@idiap.ch*

# Selected publications

All publications related to the project AMIDA are available on the page:
*http://publications.amiproject.org*

An Adaptive Initialization Method for Speaker Diarization based on Prosodic Features.
*D. Imseng and G. Friedland*
In Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Dallas, 2010.

Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis.
*O. Glembek, L. Burget, N. Dehak, N. Brümmer, and P. Kenny*
In Proc. Proc. ICASSP 2009, 2009, p. 4.

Application of Out-Of-Language Detection To Spoken-Term Detection.
*P. Motlicek and F. Valente*
In Proc. 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, 2010.

Cascaded Model Adaptation for Dialog Act Segmentation and Tagging.
*U. Guz, G. Tur, D. Hakkani-Tür, and S. Cuendet*
In Journal of Computer Speech and Language, 2010.

Catchup: A Useful Application of Time-Travel in Meetings.
*S. Tucker, A. Ramamoorthy, O. Bergman, and S. Whittaker*
In Proc. Proceedings of CSCW, 2010.

Differences in head orientation behavior for speakers and listeners: An experiment in a virtual environment.
*R. Rienks, R. Poppe, and D. Heylen*
In ACM Transactions on Applied Perception, vol. 7, iss. 1, pp. 1–13, 2010.

Face Alignment Using Boosting and Evolutionary Search.
*H. Zhang, D. Liu, Mannes Poel, and A. Nijholt*
In Proc. Proceedings of the Asian Conference on Computer Vision (ACCV) 2009, 2010, pp. 110–119.

Leveraging Speaker Diarization for Meeting Recognition from Distant Microphones.
*I. D. A. Stolcke G. Friedland*
In Proc. IEEE ICASSP, 2010.

Tuning-Robust Initialization Methods for Speaker Diarization.
*F. G. D. Imseng*
In IEEE Transactions on Audio, Speech and Language Processing, 2010.

Using Audio and Visual Cues for Speaker Diarisation Initialisation.
*G. Garau and H. Bourlard*
In Proc. International Conference on Acoustics, Speech and Signal Processing, 2010.

Cancer Stage Interpretation System.
*A. N. Nguyen, M. J. Lawley, and D. P. Hansen*
The Australian e-Health Research Centre, CSIRO ICT Centre Technical Report 09/118, 2009.

A Human Benchmark for Language Recognition.
*R. Orr and D. van Leeuwen*
In Proc. Interspeech, 2009.

A Multimedia Retrieval System Using Speech Input.
*A. Popescu-Belis, P. Poller, J. Kilgour, E. Boertjes, J. Carletta, S. Castronovo, M. Fapso, A. Nanchen, T. Wilson, J. de Wit, and M. Yazdani*
In Proc. of ICMI-MLMI 2009 (11th International Conference on Multimodal Interfaces and 6th Workshop on Machine Learning for Multimodal Interaction), 2009.

A Parallel Training Algorithm for Hierarchical Pitman-Yor Process Language Models.
*S. Huang and S. Renals*
In Proc. Proc. Interspeech'09, Brighton, UK, 2009.

Any Questions? Automatic Question Detection in Meetings.
*K. Boakye, B. Favre, and D. Hakkani-Tür*
In Proc. Proceedings of the 11th Biannual IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2009.

Audio spatialisation strategies for multitasking during teleconferences.
*S. N. Wrigley, S. Tucker, G. J. Brown, and S. Whittaker*
In Proc. Interspeech 2009, 2009, pp. 2935–2938.

Automatic Out-of-Language Detection Based on Confidence Measures Derived from LVCSR Word and Phone Lattices.
*P. Motlicek*
In Proc. 10thAnnual Conference of the International Speech Communication Association, 2009, pp. 1215–1218.

Automatic vs. human question answering over multimedia meeting recordings.
*Q. A. Le and A. Popescu-Belis*
In Proc. Proc. of 10th Annual Conference of the International Speech Communication Association, 2009.

Boosting Multi-Modal Camera Selection with Semantic Features.
*B. Hörnler, D. Arsic, B. Schuller, and G. Rigoll*
In Proc. Proc. Int. Conf. on Multimedia & Expo, ICME 2009, New York, NY, USA, 2009, pp. 1298–1301.

Brno University of Technology System for Interspeech 2009 Emotion Challenge.
*M. Kockmann, L. Burget, and J. Černocký*
In Proc. Proc. Interspeech 2009, 2009, pp. 348–351.

A Multimedia Retrieval System Using Speech Input.
*A. Popescu-Belis, P. Poller, J. Kilgour, E. Boertjes, J. Carletta, S. Castronovo, M. Fapso, A. Nanchen, T. Wilson, J. de Wit, and M. Yazdani*
In Proc. of ICMI-MLMI 2009 (11th International Conference on Multimodal Interfaces and 6th Workshop on Machine Learning for Multimodal Interaction), 2009.

BUT system for NIST 2008 speaker recognition evaluation.
*L. Burget, M. Fapso, V. Hubeika, O. Glembek, M. Karafiat, M. Kockmann, P. Matějka, P. Schwarz, and J. Černocký*
In Proc. Proc. Interspeech 2009, 2009, pp. 2335–2338.

# AMIDA Final Public Report
# Abstract

This report describes the overall methodology behind the development of meeting support technologies, and gives an overview of the main achievements of the AMI-AMIDA European projects, spanning six years of multi-disciplinary research and development. The work summarized here has been primarily carried out in the context of two EU Integrated Projects, referred to as AMI (Augmented Multiparty Interaction, January 2004-December 2006) and its follow-up project AMIDA (October 2006-December 2009), both described at http://www.amiproject.org/, including detailed scientific and business portals.

As we will see in this report, the development of such technologies, their related applications, and user-centric evaluations, require to go beyond, and integrate in a principled way, the state-of-the-art in several multi-disciplinary areas, including models of group dynamics (behavioral and social sciences), audio and visual processing and recognition, models to combine multiple modalities, the abstraction of content from multiparty meetings, and issues relating to human-computer interaction. These R&D themes are underpinned by the ongoing capture of user requirements, the development of a common infrastructure, and evaluations of the resultant systems. While meetings provide a rich case study for research, and a viable application market, many of the scientific advances that have been made within the AMI and AMIDA projects are wider than any single application domain. Each of the technologies briefly discussed in this report has broad application, for example in security, surveillance, home care monitoring, and in more natural human-computer interfaces. Progress on several fronts were necessary to develop the targeted meeting support technologies, from establishing the necessary framework for successful long-term collaborative research, through to development of multiple prototypes to access (online or offline) meeting archives (including meeting browsers and automatic content linking systems). A common hardware and software infrastructure had to be established, and within this we undertook ambitious data collection and annotation efforts. We have also developed leading edge technologies in audio, visual and multimodal processing, and in content abstraction.

# AMIDA: Final Public Report

## Contents

In recent times there has been growing research interest in the recognition and understanding of interactions between people in settings such as meetings, lectures, seminars and teleconferences. The modelling and interpretation of human-human communication scenes is a challenging scientific endeavour, requiring a broad range of research advances in areas including signal processing, speech recognition, multimodal scene analysis, discourse analysis, and multimodal retrieval. The analysis and interpretation of multiparty meetings is of scientific interest since it provides a circumscribed arena for the investigation of communication scenes, as well as underpinning a number of potentially significant applications.

Meetings play a crucial role in the generation of ideas, documents, relationships, and actions within an organization. The wealth of information exchanged in meetings is often lost, at least in part, because human note taking of meeting minutes is subjective and incomplete, capturing only a fraction of the information. Multimodal recording of meetings is an attractive alternative, but such recordings will only become really useful once it is possible to recognize, structure, index and summarize them automatically.

Since the mid-1990s a number of researchers have investigated the automatic recording, recognition and interpretation of meetings [KAHM96, RL99, YGC01, LEG02, WBM+01, MBB+03, CRP+06]. From 2004, the AMI consortium has investigated the development of technologies to enhance human collaboration in the domain of meetings. AMI is concerned with the development of algorithms, models, and prototype systems that support interaction in meetings and access to meeting-related information. Our initial research was concerned primarily with the analysis of face-to-face meetings recorded in an instrumented meeting room equipped with multiple microphones and cameras, and capturing other interaction modalities including the handwriting and data projected slides. More recently, when moving from AMI to AMIDA, we have extended the focus of our work to support meetings where some of the participants may be remote, and to provide services to operate on meetings both in realtime and on an archive.

Much of the research that we have carried out has built on a corpus of 100 hours of multimodal meeting recordings annotated at a number of different levels, outlined in Section 2.2. Some of the core work of the AMI consortium has been the development of recognizers for audio and video modalities, including gesture and action recognition and audio-visual tracking. These are briefly outlined in Section 3, which is followed by a more detailed discussion of the AMI system for automatic speech transcription of meetings, from both close-talking and distant microphones (Section 3.1.1). The output of the multimodal recognizers, in particular the automatic speech transcription, forms the basis of our work in content extraction, including topic segmentation, summarization and dialogue act recognition, discussed in Section 4. A key aspect of our work has been a focus on evaluation, both at the component and system levels (see Section 6), the latter being closely tied to the design of the AMI corpus.

# 1 Multidisciplinary challenges

## 1.1 Human-to-Human Communication

Human-to-human communication is among the most advanced and complex processes known to man. Humans are high speed, highly sensitive multimodal processors, meaning that they (we) are receiving and analyzing information from multiple simultaneous inputs in real time. And, most humans are communicating in many ways with apparently little effort.

The driving vision of the AMI and AMIDA projects were to be able to analyse, model and understand multimodal human-to-human comunication scenes, and to develop technologies that will support and even enhance human communications. Computers will process digitally captured and transmitted or archived media and perform annotations on the media such that at a later point in time, or in real time, analyses and searches can be performed.

There is a lot about human-to-human communications that has been scientifically measured and documented but much more remains to be studied. In fact, there's a lot more to learn and study about human communications than is known today. In order to study, and potentially to mimic and to improve our ability to use computers with communications, it is necessary to develop advanced algorithms able to extract explicit and implicit information present in meetings, to interpret and index it, and to develop offline and online tools to retrieve that information.

## 1.2 Why Meetings?

Developing scientific models on the basis of human interaction during meetings, and studying human behaviors in meetings, has the potential to affect many processes and to have a positive impact on businesses and public service agencies. Numerous studies confirm what we all know from personal experience: meetings dominate the way people work. According to a study conducted by MCI Worldcom in 2003, on average in 2002, a business person participated in 60 meetings per month. People meet in groups for a multitude of reasons. They interact in numerous predictable and unpredictable ways and the results of their interactions are as varied as the people who participate and the projects on which they are collaborating or communicating. By definition, no two meetings are exactly alike, unless one is a recording of another. Studies of business processes also reveal that approximately 80% of the "workload" associated with a project or process happens in preparation for a meeting. In other words, many people view the "live" meeting as a milestone or deadline by which they can pace and measure their productivity and that of their colleagues. Unfortunately, for many information managers, being in perpetual meetings has reduced their ability to prepare adequately for the next meeting, perpetuating a vicious and negative cycle.

As other business processes, meetings are going digital. Increasingly, people are using computer technology alone and in conjunction with broadband networks to support their meeting objectives prior to and during an actual meeting. E-mail is used to pass around files for people to read prior to a meeting. Collaborative workspaces in corporate networks and on the Internet offer geographically distributed collaborators a virtual repository for documents related to a project or a meeting. Electronic meeting support systems, such as interactive network-connect white boards and videoconferencing appliances, are available for the benefit of those who share the same room as well as those who are in remote locations. Computer-supported collaborative work technologies, particularly those which capture human verbal and non-verbal communications (audio and video interaction) in addition to text and graphics generated during a meeting, also promise to have a long term impact on how people will prepare for and behave during and following meetings.

What if people could quickly review the full recording or an abridged and annotated playback of a previous meeting on a topic and find specific elements when preparing for an upcoming meeting? Many of the promised benefits of meeting technologies rely on the ability for computers to automatically generate fully indexed, searchable archives as well as additional layers of value in knowledge that is otherwise inaccessible to the meeting participants, their colleagues or supervisors. In order to design technologies with the potential to "unlock" the business value contained in meetings, and in larger accumulations of multimedia meeting archives, researchers in several related fields must collaborate.

### 1.3 Research Challenges

To develop these applications, the AMI and AMIDA projects have been extending the state-of-the-art in several areas, including models of group dynamics, audio and visual processing and recognition, models to combine multiple modalities, the abstraction of content from multiparty meetings, and issues relating to human-computer interaction. These R&D themes are underpinned by the ongoing capture of user requirements, the development of a common infrastructure, and evaluations of the resultant systems. While meetings provide a rich case study for research, and a viable application market, many of the scientific advances being made within AMI and AMIDA are wider than any single application domain. Each of the above technologies has broad application, for example in security, surveillance, home care monitoring, and in more natural human-computer interfaces.

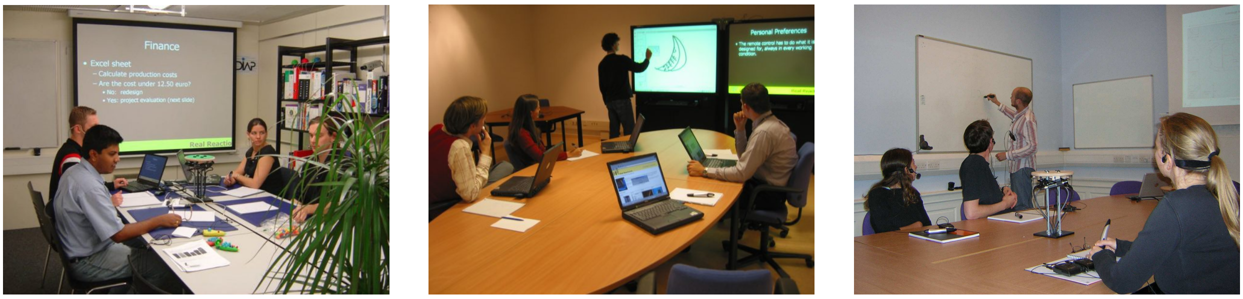


Figure 1: The three AMI- AMIDA instrumented meeting rooms at Idiap Research Institute (left), TNO (centre) and the University of Edinburgh (right).

# 2 Instrumented meeting rooms and AMI corpus

## 2.1 Instrumented meeting room

Much of our research is built on the use of instrumented meeting rooms to collect recordings of multiparty meetings. Three standardized meeting rooms were designed and constructed at AMI partners IDIAP, TNO and the University of Edinburgh (Figure 1).

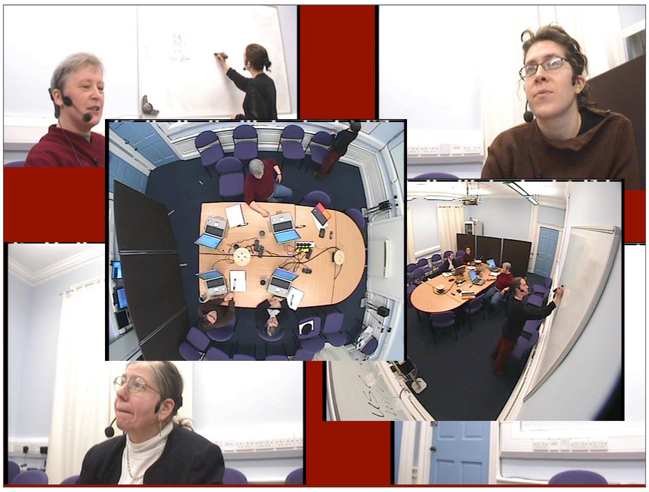


Figure 2: *Video captured from an AMI instrumented meeting room*

These *instrumented meeting rooms*, which were designed for the collection of four person meetings, all contained a set of standardized recording equipment:

- six cameras — four providing close-up views of the participants, two providing a view of the whole room;
- twelve microphones — a headset microphone per participant and an 8-element circular microphone array;
- data projector capture (VGA);
- whiteboard capture;
- digital pen capture.

The general layout of the instrument meeting rooms is also illustrated in Figure 3 There were also additional recording devices in each of the rooms, including an additional microphone, a binaural manikin and additional cameras. Figure 1 shows the three AMI meeting rooms, and Figure 2 shows stills captured from the six cameras in the University of Edinburgh instrumented meeting room during an AMI meeting. To ensure data diversity, the rooms differ in layout and wide-angle cameras. In addition to this fixed infrastructure, we have developed and worked with a number of configurations for portable equipment that can be used in any room. These range from deliberately low-cost systems to ones that use high-specification microphones and a 360 degree camera for complete coverage.

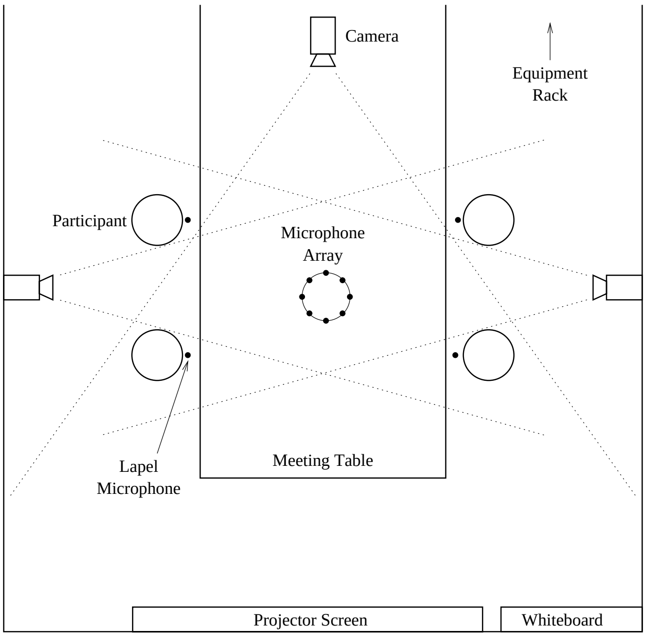


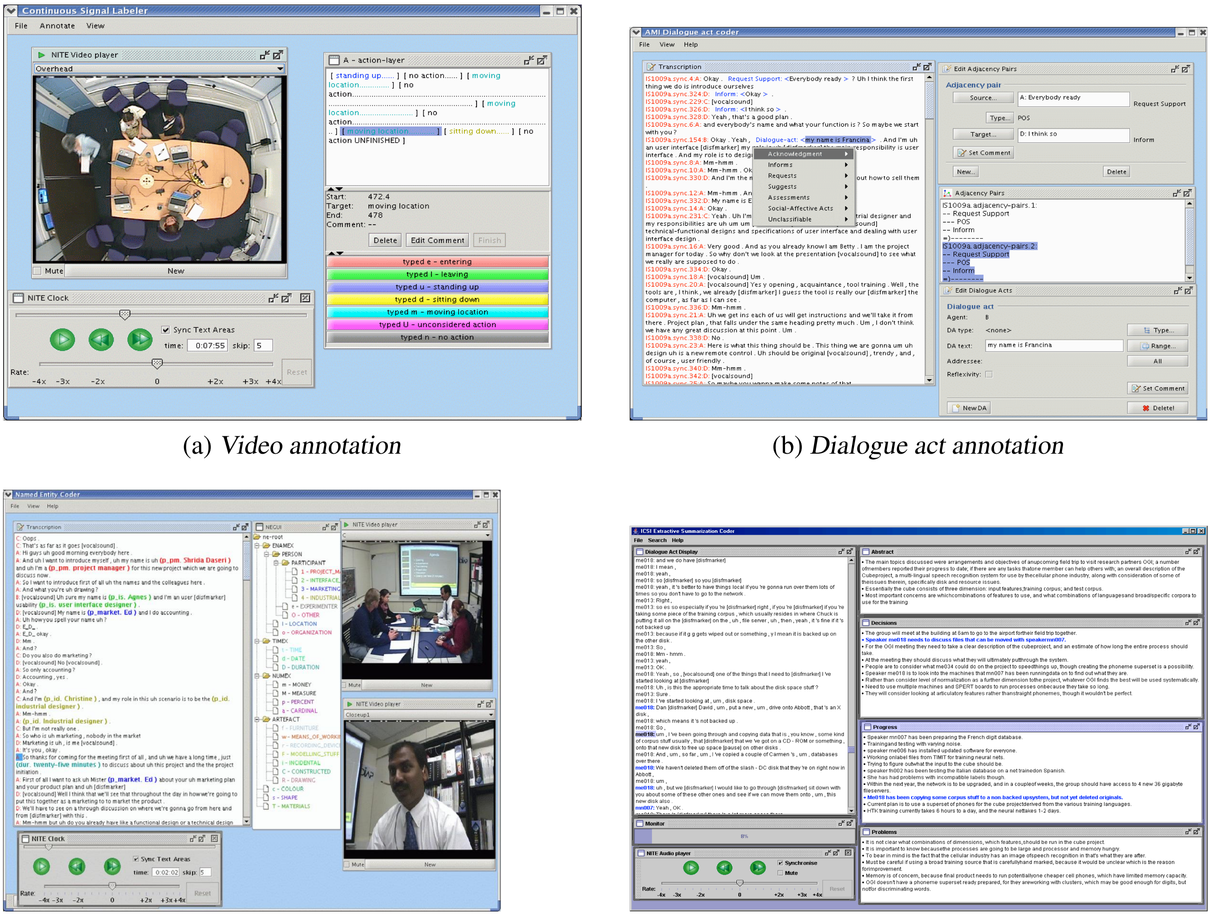
Figure 3: Typical layout of an instrument meeting room.

## 2.2 AMI corpus

These instrumented meeting rooms were used to record the AMI Meeting Corpus [Car07], which consists of 100 hours of meeting recordings, with the different recording streams synchronized to a common timeline. The corpus includes manually produced orthographic transcriptions of the speech used during the meetings, aligned at the word level. In addition to these transcriptions, the corpus includes manual annotations that describe the behaviour of meeting participants at a number of levels. These include dialogue acts, topic segmentation, extractive and abstractive summaries, named entities, limited forms of head and hand gestures, gaze direction, movement around the room, and where heads are located on the video frames. Not all 100 hours of meetings have been marked with all kinds of annotations. The linguistically motivated annotations have been applied most widely, covering at least 70% of the corpus in all cases. The annotations were carried out using NXT (the NITE XML Toolkit) [CEHK05], an open source XML-based infrastructure for the annotation and management of multimodal recordings[2]. Examples of NXT annotation tools used with the AMI corpus are shown in Figure 4.

The corpus consists of two types of meetings: a design scenario, and naturally occurring meetings in a range of domains. About 70% of the corpus was elicited using the scenario in which the participants play different roles in a design team, taking a design project from kick-off to completion over the

[2] `http://groups.inf.ed.ac.uk/nxt/`



(a) *Video annotation*

(b) *Dialogue act annotation*

(c) *Named entity annotations*

(d) *Summarization annotation*

Figure 4: *AMI corpus annotations using NXT*

course of a day. The scenario meetings consist of a series of four meetings, attended by four participants, who had tasks to accomplish between meetings. The participant roles were driven in real-time by emails and web information. There are several advantages to recording scenario meetings. First, it enabled us to control the domain, making it easier to understand the content of the meetings, and to enable the construction of deeper approaches to content extraction. Second, the construction of a meeting scenario enabled outcome measures to be defined, including preferred design outcomes. Third, the fact that participants were not part of a real organization made it much easier to understand what they knew and what motivated them. Fourth, scenario meetings are replicable, and thus enable system-level evaluations, such as the task-based evaluation discussed in Section 6.

The corpus is publicly available on the web at `http://corpus.amiproject.org`, and is released under a licence that is based on the terms of the Creative Commons Attribution NonCommercial ShareAlike 2.5 Licence.

### 2.3 User requirements

The work of the project was underpinned by a set of requirements for using multimodal recordings of meetings, captured through observation of participants in business meetings, interviews, focus groups, and questionnaires. The captured use cases were structured into two categories: for a *meeting browser* and for a *remote meeting assistant*.

The principal use case for meeting browsing is to help team members look up specific details about a meeting, or a set of meetings, whether they have attended it or not. Other browser use cases include meeting audits (e.g., to track back decisions), catching up on a meeting you are late for, and accessing contents of previous meetings while a meeting is underway.

The principal use case for remote meetings is meeting monitoring: that is, to only join a running remote meeting when an alert is supplied that the time is right. Other remote meeting use cases include assistants for live remote meetings (providing indications of the group state), and improving the experience of attending a remote meeting.

These use cases always drove the processing technologies that were developed in AMI and AMIDA.

# 3 Audio-video processing

## 3.1 General goals

AMI and AMIDA work in audio-visual processing was primarily concerned with the development of algorithms that can automatically answer each of the following questions from the raw audio-video streams:

- What has been said during the meeting? (Speech recognition)
- What acoustic events and keywords occur in the meeting? (Keyword spotting)
- Who and where are the persons in the meeting? (Localization and tracking)
- Who in the meeting is acting or speaking? (Speaker tracking)
- How do people act in the meeting? (Gesture and action recognition)
- What are the participants' emotions in the meeting? (Emotion)
- Where or what is the focus of attention in meetings? (Focus of attention)

### 3.1.1 Speech recognition

Automatic transcription of speech in meetings is of crucial importance for tasks such as meeting analysis, content analysis, analysis of dialogue structure, and summarization.

AMI and AMIDA developed systems for the two types of microphone configurations in the instrumented meeting rooms (close-talking headset microphones and tabletop microphone arrays), focusing on the headset microphone conditions to develop core acoustic modeling approaches, but with an overall orientation to tabletop microphone arrays, which are less intrusive. In particular, the AMI speech recognition effort has addressed several research issues including the following:

- Microphone array beamforming: filtering and combining the individual microphone signals to enhance signals coming from a particular location (and suppressing competing locations)
- Development of novel acoustic parameterizations, including approaches based on posterior probability estimation
- Automatic construction of domain-specific language models using text extracted from the web
- Acoustic segmentation
- Development of a flexible large vocabulary decoder, based on a weighted finite state transducer formalism
- Development of a real-time speech recognition system that can be used in applications that need instant or quick results

AMI and AMIDA have developed an evaluation framework that is generic, flexible, comparable, and that allows us to conduct research and development in a stable environment. Using this framework, our system obtains exceptionally good results on AMI meeting data; in international technology evaluations organized by NIST, no other system was significantly more accurate than the AMI system on close-talking microphones. This system has been used to decode the complete AMI corpus (using an n-fold cross-validation technique), and these transcriptions have been used for tasks such as summarization and topic segmentation.

### 3.1.2 Keyword spotting

In acoustic keyword spotting (KWS), the goal is to find keywords and their position in speech data. AMI has developed three approaches: acoustic, LVCSR, and a hybrid approach.

In the acoustic approach, a keyword score is obtained by comparing the posterior probability of the keyword phonetic model, with a background model. This is very fast since many of the key parameters may be pre-computed. It is relatively precise (the precision increases with the length of the keyword) and any word can be searched provided its phonetic form is available. It is ideal for on-line applications (such as monitoring remote meetings), but it is not suitable for browsing huge archives, as it needs to process all the acoustic data for each search.

The LVCSR lattice approach locates the keywords in lattices generated by a large vocabulary continuous speech recognition system. Given the output of the speech recognizer, this approach is very fast, but it is accurate only for frequently occurring words. There is a degradation in performance for less common words, which is a drawback, since these words (such as technical terms and proper names) carry most of the information and are likely to be searched by users. Therefore, this approach has to be complemented by a method unconstrained by the recognition vocabulary.

The hybrid phoneme lattice approach is based on the construction of graphs of phoneme probabilities, from which the phonetic form of the keyword may be extracted. This is a reasonable compromise in terms of accuracy and speed. Currently, AMI work on indexing phoneme lattices using tri-phoneme sequences is advancing and preliminary results show a good accuracy/speed trade-off for rare words.

### 3.1.3 Speaker tracking

The objective of speaker tracking is to segment, cluster and recognize the speakers in a meeting, based on their speech. The first approach developed in AMI uses the acoustic contents of the microphone signal to segment and cluster speakers. In the NIST evaluations this system produced very good results for speech activity detection (the lowest error rate reported) and for speaker diarization (”who spoke when”). The second approach developed in AMI, based on cross-correlations between microphone signals operates in real time, and has been integrated with the online keyword spotter. Real-time diarization required us to develop novel initialization methods for our diarization engine, because our performance was worse on segments shorter than 10 minutes. We also obtained performance improvements for diarization by dealing explicitly with overlapped speech and by adding video features to the more usual audio ones.

### 3.1.4 Localization and tracking

Location coordinates of each person in the meeting are an essential input to various meeting analysis tasks, including focus of attention and action recognition. The steps required are identification, localization, and tracking. For identification, generative approaches have proven to be the most robust so in AMI a variety of models with different trade offs between speed and accuracy have been used (e.g., based on Gaussian mixtures and HMMs). The algorithms have been developed as a machine vision package for the open source machine learning library, TORCH (`http://www.torch.ch`), which we extended within AMI. For localization and tracking AMI developed, applied, and evaluated four different methods including approaches based on dynamic Bayesian networks, active shape trackers using particle filters, and face trackers based on skin colour.

### 3.1.5 Gesture and action recognition

We have defined a set of actions and gestures that are relevant for meetings (e.g., hand, body, and head gestures such as pointing, writing, standing up, or nodding). Special attention has been paid to negative signals, such as a negative response to a yes-no question, usually characterized by a head shake. This kind of gesture contains important information about the decision making in meetings, but can be very subtle and involve little head movement, making automatic detection very difficult.

For gesture recognition two methods were applied: Bayesian Information Criterion and an Activity Measure approach. We extracted, for each person in the meeting, the 2D location of the head and hands, a set of nine 3D joint locations, and a set of ten joint angles. In addition we performed classification of the segmented data. Due to the temporal character of gestures we focused on different HMM methods. Gestures like standing up and important speech supporting gestures produced satisfactory results (100% and 85% recognition rate, respectively). However the results for the detection of negative signals were not significantly better than guessing. Detecting gestures such as shaking or nodding and negative signals is still a challenging problem that requires methods capable of detecting very subtle head movements.

Real-time gesture and action recognition presents additional challenges, especially for invariance to viewpoint, lighting, background, and person appearance in a video. Our approach to this problem combines example-based pose recovery with action classification based on Common Spatial Patterns (CSP).

### 3.1.6 Focus of Attention

Gaze detection requires higher resolution of facial images than what is available in the AMI corpus. As an approximation, we have developed algorithms for tracking the head and estimating its pose, based on a Bayesian filtering framework, which is then solved through sampling techniques. Results (evaluated on 8 minutes of meeting recordings involving a total of 8 people) were good, with a majority of head pan (resp. tilt) angular errors smaller than 10 (resp. 18) degrees. As expected, we found a variation of results among individuals, depending on their resemblance with people in the appearance training set.

In addition, we formulated focus of attention (FoA) as a classification task by automatically classifying FoA into one of the following categories: meeting participants, objects in the meeting room, and an “unfocused” location. Experiments using the ground truth head-pose pointing vectors resulted in frame-based classification rate of 68% and 47%, depending on the person’s position in the smart meeting room. Accuracy is lower than reported in other works, mainly because of the complexity of the scenes and number of categories. Exploiting other features/modalities (e.g speaking status) in addition to the head pose can be used to disambiguate FoA classification. We found that using the estimated head-pose instead of the ground truth did not degrade the results strongly (about 9% decrease, thus much less than the differences w.r.t. position in the meeting room), which was encouraging given the difficulty of the task. We also found that there was a large variation of recognition amongst individuals, which directly calls for adaption approaches such as Maximum A Posteriori techniques for the FoA recognition. These adaptation techniques, along with the use of multimodal observationsand techniques for classification in real-time on videos with a range of head resolutions, have been the topic of AMI- AMIDA research. Our real-time component for classifying visual focus of attention is used in our demonstrator for User Engagement and Floor Control.

# 4 Content Extraction

The extraction of content from multimodal meeting recordings is largely based on the results of the audio-video processing described above. To achieve accurate content extraction from meeting recordings, our emphasis has been on models and algorithms that combine modalities. Automatically extracted content enables meetings to be indexed and structured at a semantically richer level than is possible using the raw output of the audio-video recognizers. Much existing work in this area is concerned with the extraction of content from written language; a major focus of AMI has been the extension of textual approaches to multimodal settings, involving the use of prosodic, video and contextual features.

Our work in this area has included the development of automatic approaches to the segmentation and classification of phenomena such as dialogue acts [DR07], topics [HM06], and dominance and influence [RZGPP06], as well as abstractive and extractive summarization [MRMC06] and content-based automatic camera selection [AHHSR06]. Using the AMI corpus for all tasks, we have been able to agree on evaluation measures and procedures that allow us to compare different approaches and techniques, both internally and externally.

Here we focus on our advances in four areas: dialogue act recognition, topic segmentation, subjectivity, and summarization.

## 4.1 Dialogue act recognition

Dialogue acts (DA) are labels for utterances which roughly categorize the speaker's intention. They are useful for various purposes in a dialogue or meeting processing situation, such as part of a browser which highlights all points where a suggestion or offer was recognized. However, dialogue acts also serve as elementary units, upon which further structuring or discourse processing may be based. For example, the summarization components that we have developed are based on the dialogue act structure of a meeting.

Each dialog act in a meeting is given one of 15 labels, which fall into six major groups:

- Information exchange: giving and eliciting information;
- Possible actions: making or eliciting suggestions or offers;
- Commenting on the discussion: making or eliciting assessments and comments about understanding;
- Social acts: expressing positive or negative feelings towards individuals or the group;
- Other: a remainder class for utterances which convey an intention, but do not fit into the four previous categories;
- Backchannel, Stall and Fragment: classes for utterances without content, which allow complete segmentation of the material;

We have addressed the tasks of automatically segmenting the speech into dialogue acts, and assigning a label to each segment. The segmentation problem is non-trivial, since a single stretch of speech (with no pauses) from a speaker may comprise several dialogue acts—and conversely a single dialogue act may contain pauses.

Our approach to dialogue act recognition is based on a switching dynamic Bayesian network architecture which models a set of features related to lexical content and prosody and incorporates a weighted interpolated factored language model [DR07]. The switching DBN coordinates the recognition process by integrating all the available resources. The factored language model, which is learned from multiple conversational data corpora, is used in conjunction with additional task specific language models. In conjunction with this joint generative model, we have also employed a discriminative approach, based on conditional random fields, to perform a reclassification of the segmented DAs.

We have performed experiments using both automatic and manual transcriptions. The degradation when moving from manual transcriptions to the output of a speech recogniser is less than 10% absolute for both dialogue act classification and segmentation. Our experiments indicate that it is possible to perform automatic segmentation into DA units with a relatively low error rate. However the operations of tagging and recognition into fifteen imbalanced DA categories have a relatively high error rate, even after discriminative reclassification, indicating that this remains a challenging task. Adding subclassifiers for each of the ten most common tag confusions in the data improves over the accuracy of the previous state-of-the-art by about 1%.

### 4.2 Topic segmentation

Structuring a lengthy meeting by topic (and sub-topic) is a useful way of navigating a recorded meeting. Similar to dialogue act recognition, the aim is to infer automatically the sequential structure of the meeting; it differs in that the fundamental units (topics) are typically many minutes in duration.

Following Galley et al [GMFLJ03], we have explored two basic approaches to this task [HM06]. An unsupervised approach, LCSeg, does not require a training set of hand-marked topic boundaries, but can automatically infer topic boundaries as points where the statistics of text change significantly. An alternative supervised approach learns the topic boundaries, based on a hand-annotated training set. An advantage of the supervised approach is that it is possible to use additional features relating to prosody (e.g. pauses) and the structure of the conversation (e.g., speaker overlap). These additional features are also relatively independent of errors in the automatic speech transcription. In addition to locating topic segments, we have developed approaches to automatically generating labels for topics, based on the statistics of the automatically transcribed words that make up a topic.

If suitable training data is available (such as the AMI corpus), then it is possible to construct accurate topic segmentation systems using classifiers such as decision trees or conditional random fields. Both topic segmentation and topic labelling are relatively robust to speech recognition, with only small degradation in performance when comparing speech recognition output to hand transcriptions.

### 4.3 Subjectivity recognition and segmentation

Late in the project, we began work on the task of recognizing subjective content in meetings – mainly opinions, sentiments, agreements, disagreements, and other internal mental and emotional states that are expressed.

To recognize subjective utterances and distinguish ones with positive and negative sentiments, we combined several classifiers that use word n-grams, character n-grams, or phoneme n-grams. “Majority vote” and “maximise recall” classifiers both have advantages and disadvantages depending on the task and data, but “maximise recall” not only outperforms all classifiers very clearly on recall scores, but it gives better results for separating subjective vs. non-subjective utterances over ASR output.

We also developed an automatic system for detecting agreement and disagreement in meetings, and for detecting the speaker that is the target of agreements or disagreements. The system uses a combination of high-precision rules and machine learning classifers [GW09], with lexical, prosodic, dialogue act, and structural features.

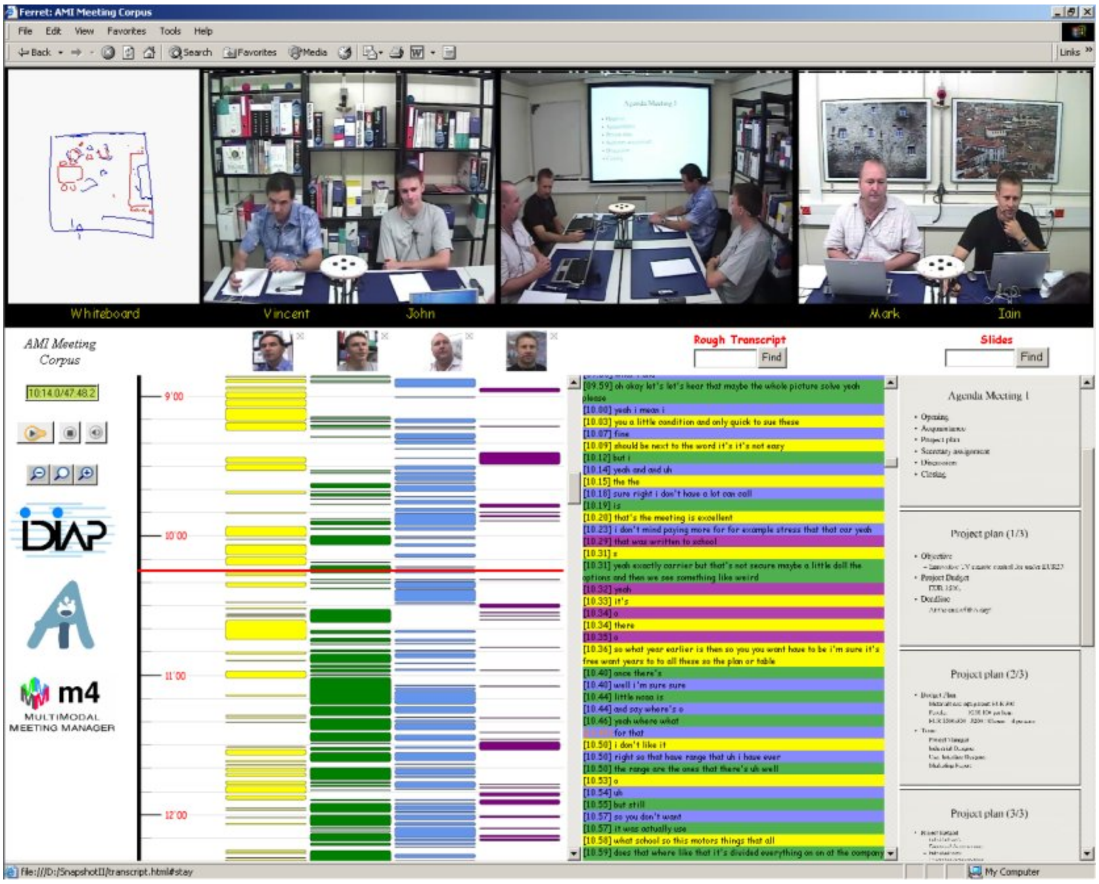
Speaker subjectivity also plays a role when analyzing conversations in terms of communicative activities. We investigated the usefulness of features that represent participant subjectivity and participant involvement (role with respect to narrated content) for the intentional segmentation of dialogue. The proposed segmentation method [NM09] outperformed a state-of-the-art one based on noun phrase coreference as detected automatically.

### 4.4 Summarization

The automatic generation of summaries provides a natural way to succinctly describe the content of a meeting, and is a very natural way for users to obtain information. In AMI we have investigated two distinct ways of constructing summaries of a meeting. *Extractive* techniques construct summaries by locating the most relevant parts of a meeting and concatenating them together to provide a 'cut-and-paste' summary, which may be textual or multimodal. *Abstractive* summaries, on the other hand, are similar to what a human summarizer might construct, generating new text to succinctly describe the meeting. Abstractive summarization is more challenging than extractive summarization, and requires relatively deep domain knowledge.

Our approach to extractive summarization is based on automatically extracting relevant dialogue acts from a meeting, as described in [MRMC06]. It thus requires (as a minimum) the automatic speech transcription and dialogue act segmentation modules described above. Lexical information is clearly extremely important for this task, but we have found it beneficial to augment information derived from the transcription with speaker features (relating to activity, dominance and overlap), structural features (the length and position of dialogue acts), prosody, and discourse cues (phrases which signal likely relevance). All these features are important to develop accurate methods for extractive summarization. Furthermore we have explored reduced dimension representations of text, based on latent semantic analysis, which also add precision to the summarization. Using an evaluation measure referred to as weighted precision, we have discovered that it is possible to reliably extract the most relevant dialogue acts, even in the presence of speech recognition errors.

We have explored "dialogue act compression", in which the extracted dialogue acts are themselves condensed, by removing irrelevant portions [MR06]. Again, taking account of speech features such as the overall intonation contour of the dialogue act helps to improve the overall performance.



[09.59] oh okay let's let's hear that maybe the whole picture solve yeah please
[10.00] yeah i mean i
[10.03] you a little condition and only quick to sue these
[10.07] fine
[10.09] should be next to the word it's it's not easy
[10.12] but i
[10.14] yeah and and uh
[10.15] the the
[10.18] sure right i don't have a lot can call
[10.19] is
[10.20] that's the meeting is excellent
[10.23] i don't mind paying more for for example stress that that car yeah
[10.29] that was written to school
[10.31] s
[10.31] yeah exactly carrier but that's not secure maybe a little doll the options and then we see something like weird
[10.32] yeah
[10.33] it's
[10.34] a
[10.34] there
[10.35] a
[10.36] so what year earlier is then so you you want have to be i'm sure it's free want years to to all these so the plan or table
[10.40] once there's
[10.40] well i'm sure sure
[10.44] little noaa is
[10.44] and say where's a
[10.46] yeah where what
for that
[10.50] i don't like it
[10.50] right so that have range that uh i have ever
[10.50] the range are the ones that there's uh well
[10.53] a
[10.54] uh
[10.55] but still
[10.57] so you don't want
[10.57] it was actually use
[10.58] what school so this motors things that all
[10.59] does that where like that it's divided everything on on at the company

Figure 5: *The AMI meeting browser JFerret*

# 5 Application prototypes

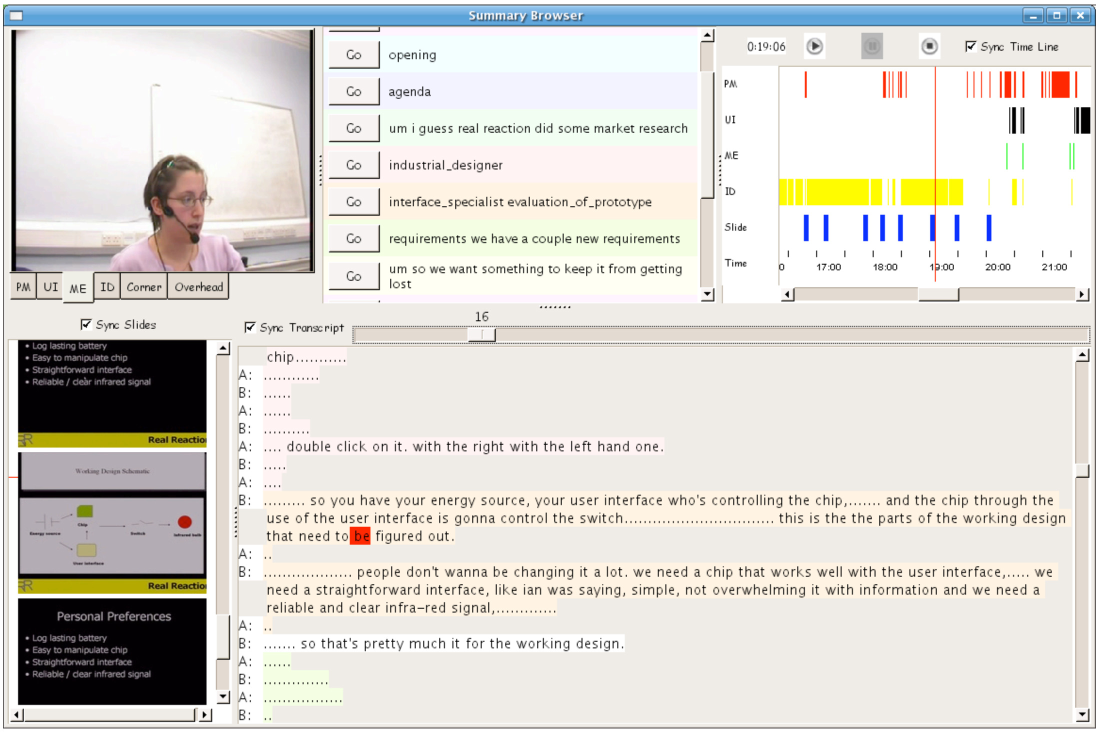
## 5.1 Meeting browsers

Figure 5 shows an example JFerret configuration, enabling browsing via keyword search on the speech-recognized transcript, search within captured slides, and browsing by speaker activity. Time-synchronized recordings that may be browsed include multiple video and audio streams and whiteboard capture. Figure 6 is an example of a JFerret browser incorporating extractive summaries and labelled topic segments, all extracted automatically. In this case the degree of compression in the textual summary is controlled by a slider. Figure 7 shows a JFerret browser constructed for browsing using dominance and influence relations between participants. Other semantically rich browser components that have been constructed include direct keyword-spotting, video hot spots, and argumentation.

Evaluation of meeting browsing has been an area in which we have made significant advances, and this is briefly discussed in Section 6.

## 5.2 Automatic Content Linking Device

The Automatic Content Linking Device (ACLD) monitors the activities of one or more users, especially their verbal output, and retrieves from a given repository, from time to time, documents that might be relevant to them. Several repositories can be searched, including, past meeting recordings, related documents, slides, minutes, etc., as well as websites. The ACLD provides just-in-time or



Go opening

Go agenda

Go um i guess real reaction did some market research

Go industrial_designer

Go interface_specialist evaluation_of_prototype

Go requirements we have a couple new requirements

Go um so we want something to keep it from getting lost

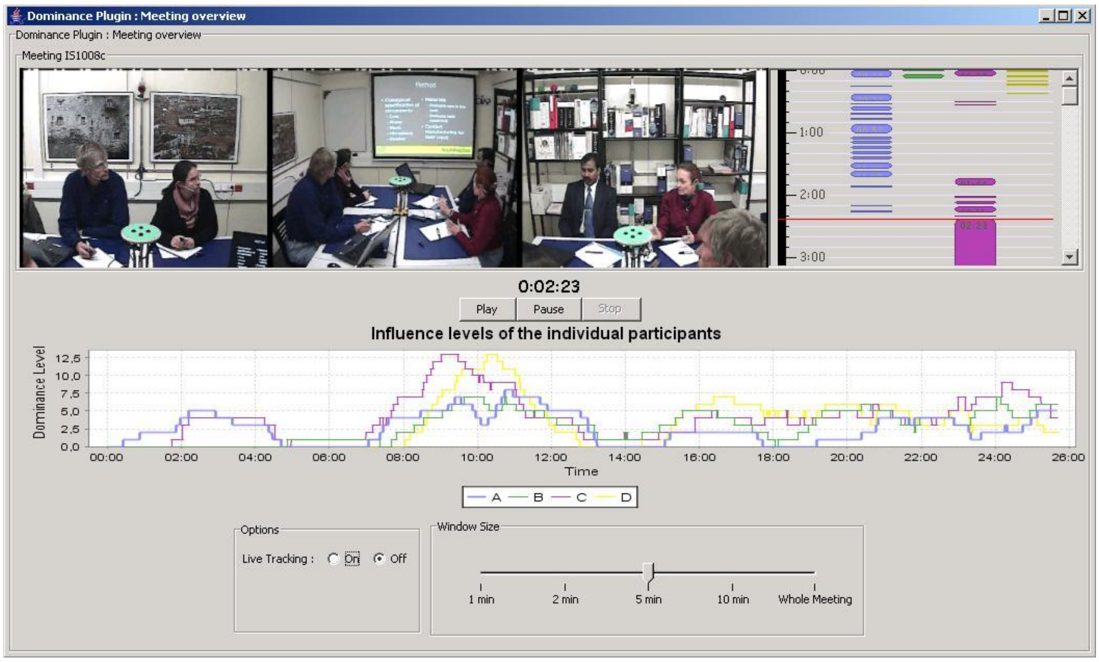
- Log lasting battery
- Easy to manipulate chip
- Straightforward interface
- Reliable / clear infrared signal

Personal Preferences

- Log lasting battery
- Easy to manipulate chip
- Straightforward interface
- Reliable / clear infrared signal

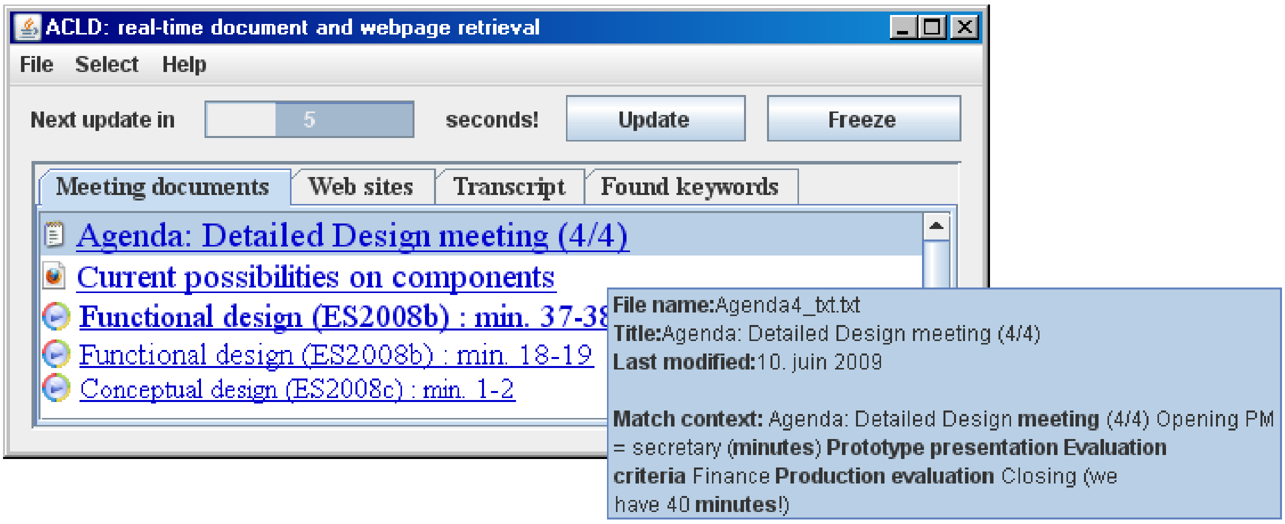
chip...........
A: ............
B: ......
A: ......
B: ..........
A: .... double click on it. with the right with the left hand one.
B: .....
A: ....
B: ......... so you have your energy source, your user interface who's controlling the chip,....... and the chip through the use of the user interface is gonna control the switch.............................. this is the the parts of the working design that need to be figured out.
A: ..
B: ................. people don't wanna be changing it a lot. we need a chip that works well with the user interface,..... we need a straightforward interface, like ian was saying, simple, not overwhelming it with information and we need a reliable and clear infra-red signal,............
A: ..
B: ....... so that's pretty much it for the working design.
A: ......
B: .............
A: ................
B: ..

Figure 6: *A configuration of the JFerret browser designed to enable browsing via automatically identified topic segments and summarization.*



Live Tracking : ○ On ● Off

Figure 7: *A configuration of the JFerret browser designed to enable browsing via dominance and influence relations.*



Agenda: Detailed Design meeting (4/4)

Current possibilities on components

Functional design (ES2008b) : min. 37-38

Functional design (ES2008b) : min. 18-19

Conceptual design (ES2008c) : min. 1-2

**File name:**Agenda4_txt.txt
**Title:**Agenda: Detailed Design meeting (4/4)
**Last modified:**10. juin 2009

**Match context:** Agenda: Detailed Design **meeting** (4/4) Opening PM = secretary (**minutes**) **Prototype presentation Evaluation criteria** Finance **Production evaluation** Closing (we have 40 **minutes**!)

Figure 8: *The ACLD UI with four superposed tabs, showing the list of relevant documents at a given moment, with explicit labels. Hovering over a label displays the metadata associated with the document, as well as excerpts where the keywords were found, with keywords in boldface.*

query-free access to these multimedia documents, based on speech from ongoing discussions. As participants in meetings often mention such documents, which contain facts that are currently discussed, but do not usually have the time to search for them, the ACLD aims at retrieving and presenting the documents automatically, hence the idea of content linking between documents and discussions. The ACLD is the first such system that is fully implemented in a multimodal interaction context.

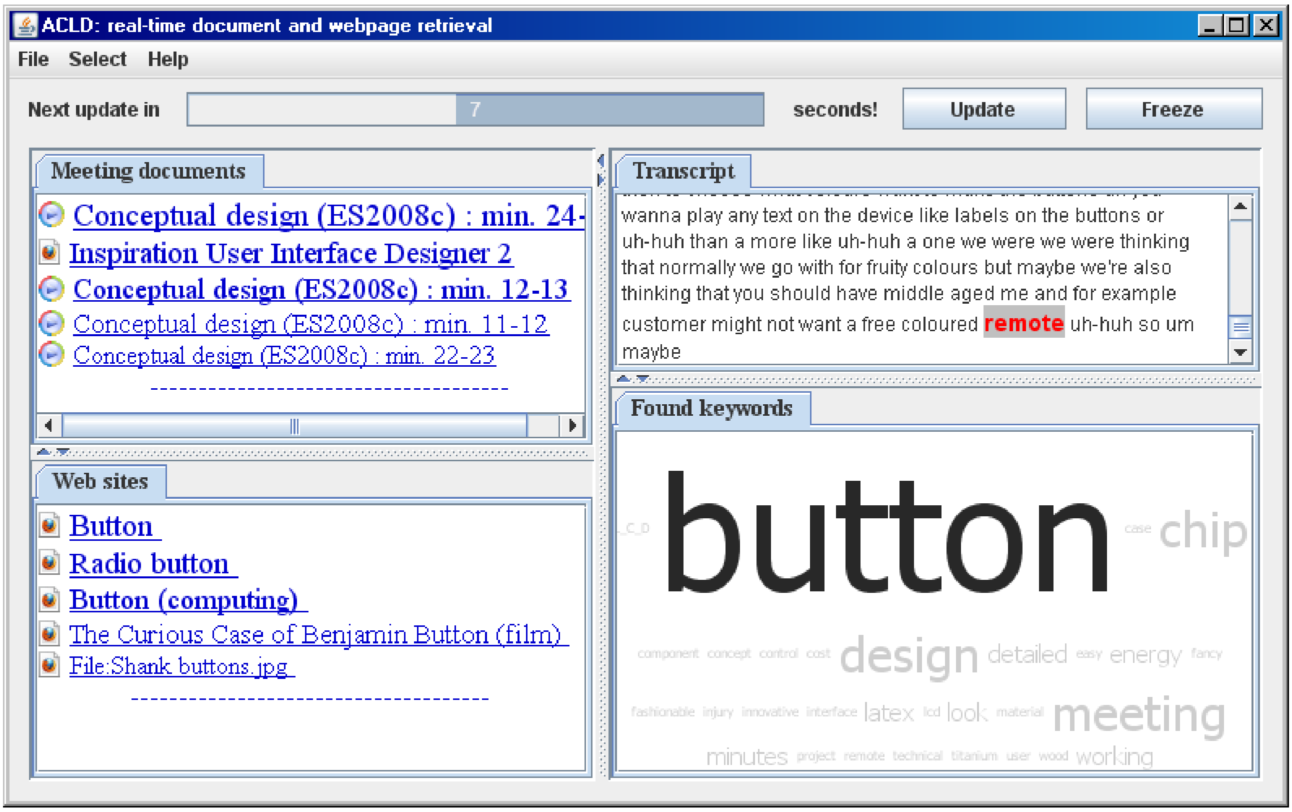
The architecture of the ACLD includes the following modules: the Document Bank Creator and Indexer; the Query Aggregator, which submits queries at regular and frequent intervals during a meeting; the User Interface, which displays the current search results, as “clickable” document links, and provides through these links access to past documents, meeting recordings, and web pages.

The ACLD is intended to be used during meetings, but the system can also be demonstrated or tested even when there is no meeting happening, by replaying a group’s meeting from an archive such as the AMI Meeting Corpus as a live meeting, and building a repository from the group’s previous meetings and associated documents. Figure 8 shows a snapshot of the Modular UI with all widgets superposed as tabs, while Figure 9 shows the Modular UI with all four widgets represented as separate tabs, occupying the four quarters of the window.

## 5.3 User Engagement and Floor Control

Remote participants in hybrid meetings often have problems to follow what is going on in the (physical) meeting room they are connected with. The User Engagement and Floor Control (UEFC) is a video conferencing system for participation in hybrid meetings [odAHH$^+$09]. The system uses modules for online speech recognition, real-time visual focus of attention as well as a module that signals who is being addressed by the speaker. A built-in keyword spotter allows an automatic meeting assistant to call the remote participant’s attention when a topic of interest is raised, pointing at the transcription of the fragment to help him catch-up.

The system demonstrates how these recognition and generation modules could be used to support remote meeting participation and make remote meetings more engaging by giving remote participants more control in discussions and decision-making processes. The UEFC system is intended for a



Conceptual design (ES2008c) : min. 24-

Inspiration User Interface Designer 2

Conceptual design (ES2008c) : min. 12-13

Conceptual design (ES2008c) : min. 11-12

Conceptual design (ES2008c) : min. 22-23

Button

Radio button

Button (computing)

The Curious Case of Benjamin Button (film)

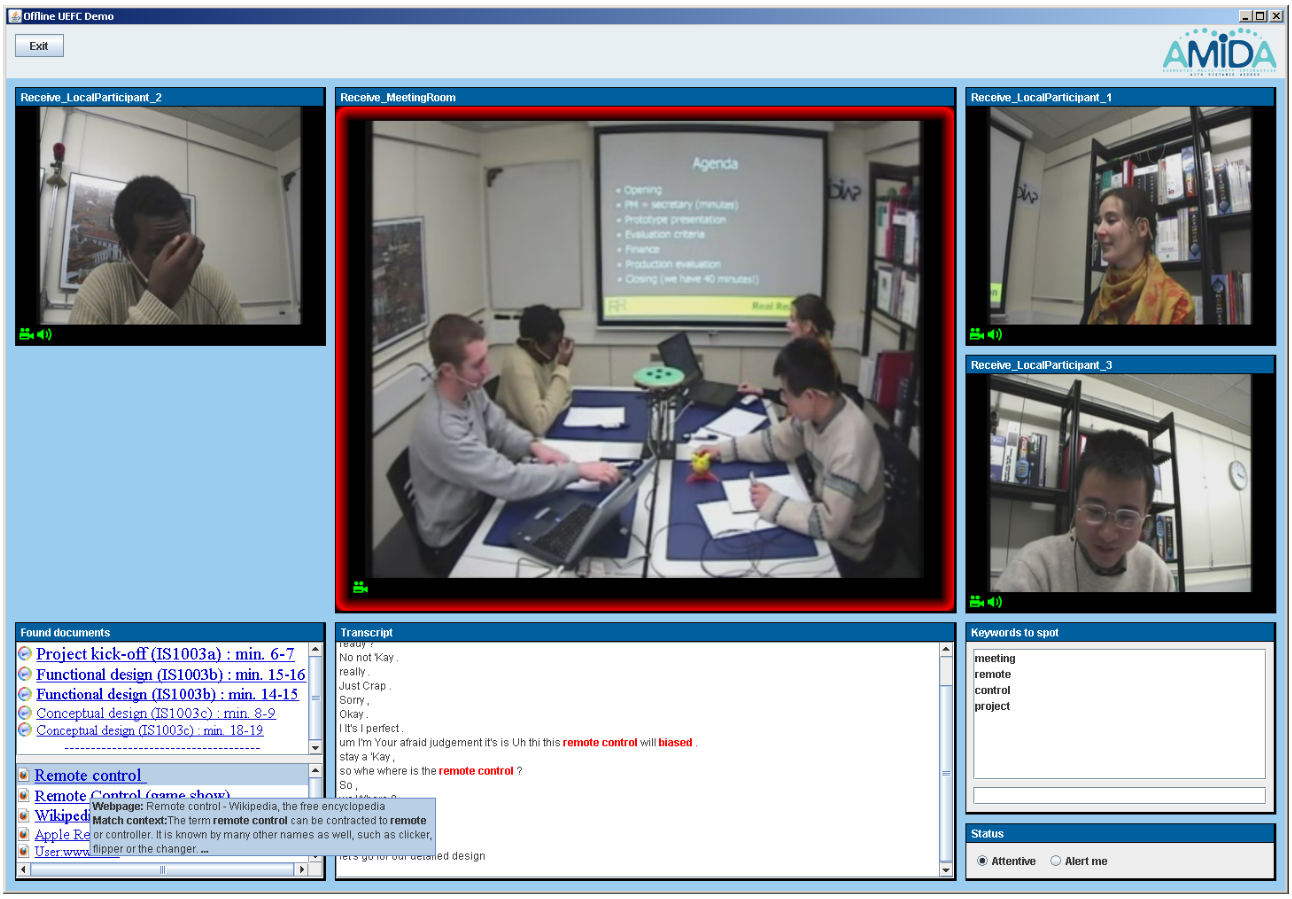
File:Shank buttons.jpg

wanna play any text on the device like labels on the buttons or uh-huh than a more like uh-huh a one we were we were thinking that normally we go with for fruity colours but maybe we're also thinking that you should have middle aged me and for example customer might not want a free coloured **remote** uh-huh so um maybe

Figure 9: *The ACLD Modular UI with the four widgets displayed side-by-side (wide screen mode).*

remote participant that has a fast internet connection and a desktop computer screen, unlike a previous mobile assistant. The use case scenario is that of a participant that cannot or has chosen not to attend the meeting continuously, who may have a special role in a project group and who is interested in some agenda items more than in others and who will either devote ”continuous partial attention” to the meeting or is multi-tasking.

Figure 10 shows the current GUI interface of the integrated system.



Project kick-off (IS1003a) : min. 6-7
Functional design (IS1003b) : min. 15-16
Functional design (IS1003b) : min. 14-15
Conceptual design (IS1003c) : min. 8-9
Conceptual design (IS1003c) : min. 18-19

Remote control
Remote Control (game show)
Wikipedi
Apple Re
User:www

**Webpage:** Remote control - Wikipedia, the free encyclopedia
**Match context:**The term **remote control** can be contracted to **remote** or controller. It is known by many other names as well, such as clicker, flipper or the changer. **...**

ready ?
No not 'Kay .
really .
Just Crap .
Sorry ,
Okay .
I It's I perfect .
um I'm Your afraid judgement it's is Uh thi this **remote control** will **biased** .
stay a 'Kay ,
so whe where is the **remote control** ?
So ,
let's go for our detailed design

meeting
remote
control
project

Figure 10: *GUI of the UEFC demonstrator, shown here over a pre-recorded meeting. The left lower frame contains links to documents  retrieved by the document retrieval system of the integrated ACLD˜-ĩn which key phrases occur that the user has specified in the right lower frame.*

# 6 Evaluation

Evaluation of technology underpins all our research in AMI and AMIDA. Evaluation protocols are generally accepted to be a strong driving force for progress, since they enable researchers to measure their research outputs with some degree of objectivity, and—through the development of joint evaluation protocols—it becomes possible to calibrate research outputs against other laboratories or projects. Furthermore, for annotated corpora to become truly useful to the research community corresponding evaluation protocols are required. We have performed evaluation both at the component technology level and at the system level, and the AMI corpus was designed to support evaluation at both levels.

At the component level, in addition to internal evaluations in a common setting, we have participated in—and contributed data to—the the NIST Meeting Recognition (RT) evaluations[3] and the CLEAR evaluations[4] of focus of attention and face detection. Additionally, the AMI corpus, together with automatic speech recognition output, was provided to the Cross Language Evaluation Forum[5] (CLEF) for their 2007 evaluation on cross-lingual question answering.

Collaborative evaluation protocols are under development for a number of areas including dominance relations, speech summarization, dialogue act segmentation and tagging. These tasks are harder to evaluate compared with recognition tasks with an unambiguous ground truth, and there are several research challenges to address in developing these evaluations, relating to high inter-annotator disagreement, and the need for subjective human judgements.

Content extraction tasks, such as summarization or topic segmentation, are somewhat artificial as a stand-alone task, and are often carried out within some other context (such as browsing). In such cases, *extrinsic evaluation* approaches may be preferred, in which a task is evaluated in the context of a larger scenario, such as a meeting browser. In AMI we have developed a framework for extrinsic evaluation of browser components, that we call the *Browser Evaluation Test* (BET) [WFTW05]. The BET provides a framework for the comparison of arbitrary meeting browser setups, where setups differ in terms of which content extraction or abstraction components are employed. The BET consists of a set of experiments in which test subjects have to answer true/false questions about *observations of interest* for a meeting recording. The test subject uses the browser under test to answer these questions, given a time limit (typically half the meeting length).

Finally, we have also developed a task-based evaluation [Piv07] that is supported by the design of the AMI corpus. As outlined above, about 70% of corpus meetings are based on a replicable design team scenario. In the current version of the task-based evaluation, a new team takes over for the fourth meeting, with access to the previous three meetings. The evaluation compares team performance in the existing case with basic meeting records (including powerpoint files, emails and minutes), with a basic AMI meeting browser, and with a task-based browser (figure 11). The task-based evaluation is in terms of both objective measures such as design quality, meeting duration, assessment of outcome, and behaviourial measures of leadership, and subjective measures including browser usability, workload (mental effort), and group process.

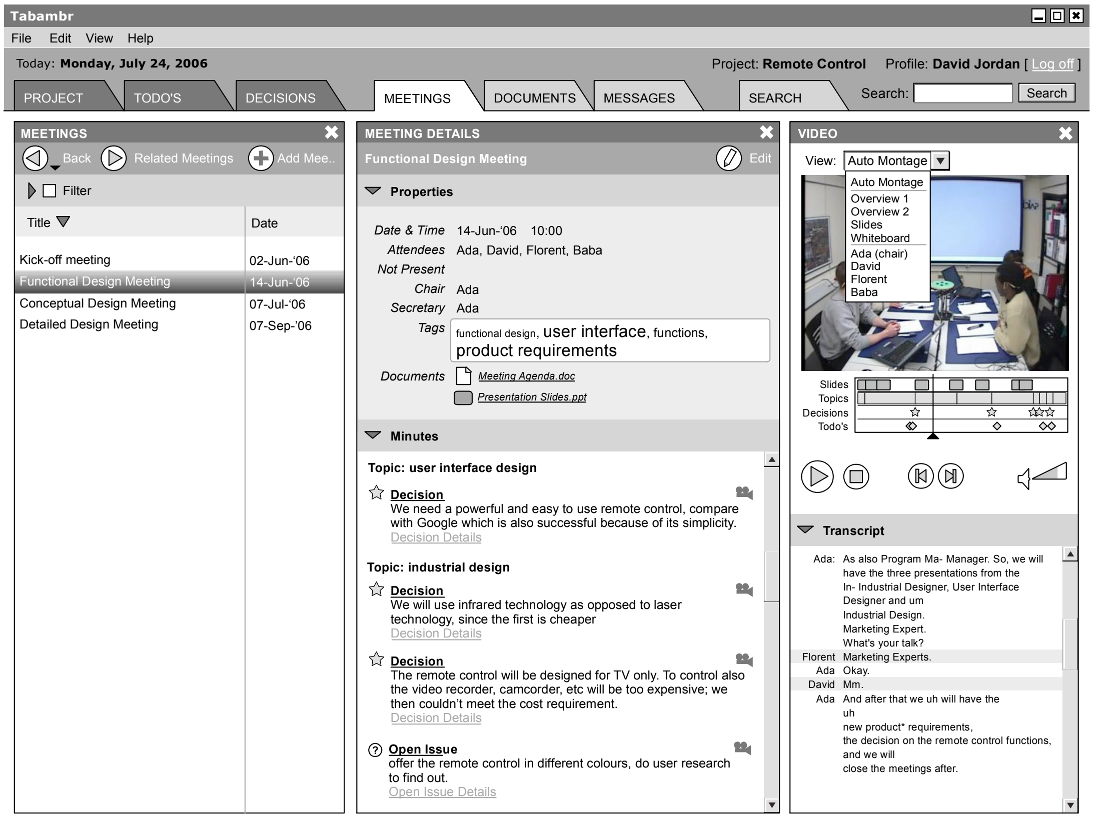
# 7 Exploiting our results

The techniques that we have used to develop our technologies rely heavily on example data, so one important question to ask is whether or not they will still work when they are used on new data that

[3] http://www.nist.gov/speech/tests/rt/

[4] http://www.clear-evaluation.org/

[5] http://www.clef-campaign.org/



Today: **Monday, July 24, 2006**

Project: **Remote Control** Profile: **David Jordan** [ Log off ]

| Title | Date |
|---|---|
| Kick-off meeting | 02-Jun-'06 |
| Functional Design Meeting | 14-Jun-'06 |
| Conceptual Design Meeting | 07-Jul-'06 |
| Detailed Design Meeting | 07-Sep-'06 |

**Functional Design Meeting**

| | |
|---|---|
| *Date & Time* | 14-Jun-'06 10:00 |
| *Attendees* | Ada, David, Florent, Baba |
| *Not Present* | |
| *Chair* | Ada |
| *Secretary* | Ada |
| *Tags* | functional design, user interface, functions, product requirements |
| *Documents* | *Meeting Agenda.doc*<br>*Presentation Slides.ppt* |

**Topic: user interface design**

☆ **Decision**
We need a powerful and easy to use remote control, compare with Google which is also successful because of its simplicity.
Decision Details

**Topic: industrial design**

☆ **Decision**
We will use infrared technology as opposed to laser technology, since the first is cheaper
Decision Details

☆ **Decision**
The remote control will be designed for TV only. To control also the video recorder, camcorder, etc will be too expensive; we then couldn't meet the cost requirement.
Decision Details

? **Open Issue**
offer the remote control in different colours, do user research to find out.
Open Issue Details

| | |
|---|---|
| Ada: | As also Program Ma- Manager. So, we will have the three presentations from the In- Industrial Designer, User Interface Designer and um Industrial Design. Marketing Expert. What's your talk? |
| Florent | Marketing Experts. |
| Ada | Okay. |
| David | Mm. |
| Ada | And after that we uh will have the uh new product* requirements, the decision on the remote control functions, and we will close the meetings after. |

Figure 11: *Task-based browser used in the task-based evaluation protocol*

comes from different sources — meeting rooms with different microphones and different acoustics, or different kinds of meetings, for instance. There are two ways in which we have addressed this important issue during the final stages of AMIDA. The first is by working closely with a new partner, Noldus Information Technology, to try some of our components in new settings that match requirements they have, and the second is by trying "mini-projects" with a range of interested companies that test our work on their data.

Noldus' flagship product, The Observer XT, is a software package for the collection and analysis of behavioural data. We modified The Observer XT to include a mechanism that can start any external process that creates ASCII or XML event logs based on the media being observed and reads those event logs back in as automatic annotations. We then used this mechanism to attach two kinds of AMIDA processing to The Observer — gesture recognition, reconceived as activities related to eating and drinking in Noldus' "Restaurant of the Future", and "comic strip" summaries. Although we cannot formally evaluate the results in the same way as we can browsers operating on our scenario meetings, the gesture recognition is correct 83-92% of the time, depending on the activity. The summarization software integrated as expected, with results that look compelling. In addition to these activities, we designed a route by which real-time speech recognition results could be fed back into The Observer XT, providing automatic transcription for the data analysts that they would find valuable.

Our mini-projects, joint with nine companies ranging from the very small to the very large, tested several of our components and ideas, including summarization, content linking, and user engagement, on meeting data of interest to them. We expect several of our mini-projects to result in longer-term collaboration between AMIDA partners and industries. As with our collaboration with Noldus, formal

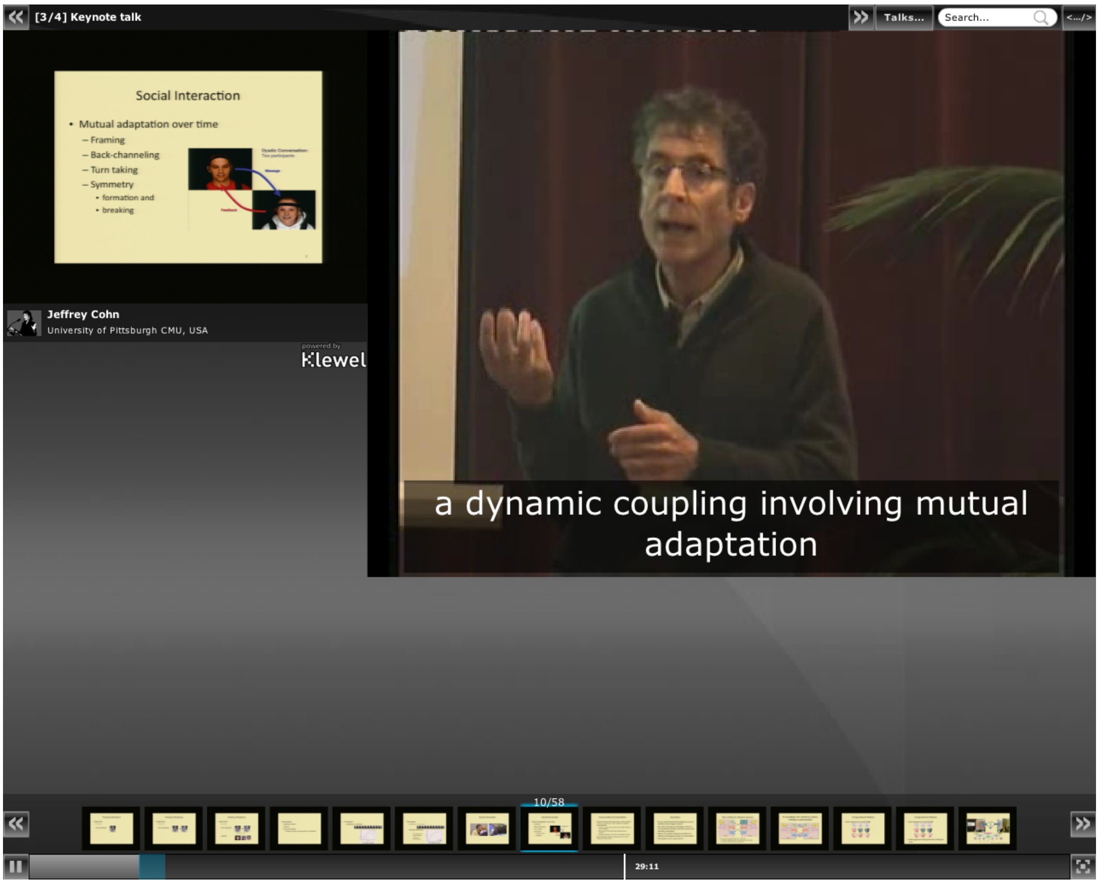


Figure 12: *Klewel-AMIDA mini-project demonstration.*

evaluation of these results is not possible, but this work has allowed to to demonstrate that we can integrate our components with wider systems and produce meeting support technologies in which companies are actively interested. Joanne Celens, CEO of Synthetron, says of our mini-project programme,

> Being a SME, following AMI since a few years, we are very enthusiastic with the mini-project experience... We were able to translate [AMIDAs latest research results] into a practical application with AMIDA researchers in very short time, testing the first "comics format" reports with several of our end clients in less then a month allowing pragmatic and quick cycle time.

Figure 12 shows the result of enhancing Klewel's lecture access system with AMIDA speech recognition technology. Including speech recognition allows Klewel to both close caption the lecture and to give the users the ability to search not just the slides, but also the speech transcription. The speech recognition is extremely accurate on this demonstration, but the interface enhancements still look useful when the quality is somewhat lower.

This demonstration is available on the internet at `http://www.klewel.com/amida_asr_demo/`.

# 8 Training

Over the course of the AMI and AMIDA projects, we provided training for young researchers. As well as supporting relevant summer schools and workshops, and the Euromasters scheme in Language and Speech (`http://www.cstr.ed.ac.uk/emasters/`), we hosted around ninety researchers at all levels from undergraduate to post-doctoral fellows for what were usually 3-12 month placements at our labs. The researchers represented a wide spread of European and non-European nationalities, working in all of the disciplines that our collaboration brings together. The trainees' comments were overwhelmingly positive about their experience, with many of the placements leading us recognize new angles to our research problems.

## References

[AHHSR06] M. Al-Hames, B. Hörnler, C. Scheuermann, and G. Rigoll. Using audio, visual, and lexical features in a multi-modal virtual meeting director. In *Proc. MLMI '06*, 2006. AMI-164.

[Car07] J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation Journal*, 41(2):181–190, 2007.

[CEHK05] J. Carletta, S. Evert, U. Heid, and J. Kilgour. The NITE XML toolkit: data model and query. *Language Resources and Evaluation Journal*, 39(4):313–334, 2005.

[CRP$^+$06] L. Chen, R.T. Rose, F. Parrill, X. Han, J. Tu, Z. Huang, M. Harper, F. Quek, D. McNeill, R. Tuttle, et al. VACE multimodal meeting corpus. *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2006.

[DR07] Alfred Dielmann and Steve Renals. DBN based joint dialogue act recognition of multiparty meetings. In *Proc IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP '07)*, 2007.

[GMFLJ03] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In *Proc ACL '03*, pages 21–26, 2003.

[GW09] Sebastian Germesin and Theresa Wilson. Agreement detection in multiparty conversation. In *Proceedings of ICMI-MLMI 2009*, pages 7–14, Cambridge, MA, 2009.

[HM06] P.-Y. Hsueh and J. Moore. Automatic topic segmentation and labeling in multiparty dialogue. In *Proc IEEE/ACL SLT '06*, 2006. AMI-203.

[KAHM96] R. Kazman, R. Al Halimi, William Hunt, and Marilyn Mantei. Four paradigms for indexing video conferences. *IEEE Multimedia*, 3(1), 1996.

[LEG02] D. Lee, B. Erol, and J. Graham. Portable meeting recorder. *ACM Multimedia*, December 2002.

[MBB$^+$03] N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. Meetings about meetings: research at ICSI on speech in multiparty conversations. *in Proc. IEEE ICASSP*, 2003.

[MR06] G. Murray and S. Renals. Dialogue act compression via pitch contour preservation. In *Proc. Interspeech '06*, 2006.

[MRMC06] G. Murray, S. Renals, J. Moore, and J. Carletta. Incorporating speaker and discourse features into speech summarization. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 367–374, 2006.

[NM09] John Niekrasz and Johanna Moore. Participant subjectivity and involvement as a basis for discourse segmentation. In *Proceedings of SIGDial 2009*, pages 54–61, London, 2009.

[odAHH+09] H. J. A. op den Akker, D. H. W. Hofs, G. H. W. Hondorp, H. op den Akker, J. Zwiers, and A Nijholt. Supporting engagement and floor control in hybrid meetings. In *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions*, volume 5641 of *Springer Lecture Notes in Computer Science*. Springer, 2009.

[Piv07] W. Post, M. A. A. in 't Veld, and S. A. A. van den Boogaard. Evaluating meeting support tools. *Personal and Ubiquitous Computing*, 2007.

[RL99] D. M. Roy and S. Luz. Audio meeting history tool: Interactive graphical user-support for virtual audio meetings. In *Proc. ESCA Workshop on Accessing Information in Spoken Audio*, pages 107–110, 1999.

[RZGPP06] R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post. Detection and application of influence rankings in small group meetings. *Proc ICMI '06*, 2006. AMI-192.

[WBM+01] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner. Advances in automatic meeting record creation and access. *in Proc. IEEE ICASSP*, May 2001.

[WFTW05] P. Wellner, M. Flynn, S. Tucker, and S. Whittaker. A meeting browser evaluation test. In *Proc. ACM CHI '05*, pages 2021–2024, 2005.

[YGC01] R. Yong, A. Gupta, and J. Cadiz. Viewing meetings captured by an omni-directional camera. *ACM Transactions on Computing Human Interaction*, March 2001.