The AMI Meeting Corpus

I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner.
The AMI Project Consortium, www.amiproject.org, J.Carletta@edinburgh.ac.uk

Abstract
To support multi-disciplinary research in the AMI (Augmented Multi-party Interaction) project, a 100 hour corpus of meetings is being collected. This corpus is being recorded in several instrumented rooms equipped with a variety of microphones, video cameras, electronic pens, presentation slide capture and white-board capture devices. As well as real meetings, the corpus contains a significant proportion of scenario-driven meetings, which have been designed to elicit a rich range of realistic behaviors. To facilitate research, the raw data are being annotated at a number of levels including speech transcriptions, dialogue acts and summaries. The corpus is being distributed using a web server designed to allow convenient browsing and download of multimedia content and associated annotations. This article first overviews AMI research themes, then discusses corpus design, as well as data collection, annotation and distribution.

Keywords
Meetings, multimodal, corpus, annotation.

1 Introduction
The AMI (Augmented Multi-party Interaction) project is concerned with the development of technology to support human interaction in meetings, and to provide better structure to the way meetings are run and documented. The project has a number of instrumented meeting rooms that enable the collection of multimodal meeting recordings. For each meeting, audio (coming from multiple microphones, including microphone arrays), video (coming from multiple cameras), slides (captured from the data projector), and textual information (coming from associated papers, captured handwritten notes and the whiteboard) are recorded and time-synchronised. All of these streams are then available to be structured, browsed and queried within an easily accessible archive.

AMI is particularly concerned with the development of meeting browsers and remote meeting assistants, and the component technologies needed to achieve these demonstrators. A meeting browser is a system that enables a user to navigate around an archive of meetings, efficiently viewing and accessing the full multimodal content, based on automatic annotation, structuring and indexing of those information streams. For example, navigation may be enabled using automatic annotations such as speech transcription, identification of participants, and identification of their focus of attention at a particular time. The concept of the meeting browser may also be extended to a remote meeting assistant which will perform such operations in real time during a meeting, and enable remote participants to have a much richer interaction with the meeting.

The development of such meeting browsers depends on a number of technological advances. AMI is 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.

As part of the development process, the project is collecting a corpus of 100 hours of meetings using instrumentation that yields high quality, synchronized multi-modal recording, with, for technical reasons, a focus on groups of four people. All meetings are in English, but a large proportion of the speakers are non-native English speakers, providing a higher degree of variability in speech patterns than in many corpora. The corpus aims to benefit a range of research communities, including those working on speech, language, gesture, information retrieval, and tracking, as well as organizational psychologists interested in how groups of individuals work together as a team. This article lists AMI research themes, then describes the design, collection and annotation of the AMI Meeting Corpus.

2 AMI Research Themes
AMI research is structured according to the following themes:

1. *Definition and analysis of meeting scenarios*: To study the type of group, the nature of their interactions, and the means by which their members communicate.
2. *Infrastructure design, data collection and annotation*: To design and install infrastructure to collect data suitable for research of AMI technologies within the context of the defined scenarios.
3. *Processing and analysis of raw multi-modal data*: To research and develop techniques for the processing and analysis of audio, visual, and multimodal data streams from meetings. Specifically, development addresses the following core problems: 1) recognising what is said by participants, 2) recognising what is done by participants (physical actions), 3) recognising where each participant is, at each time, 4) recognising participants' emotional states, 5) tracking what (person, object, or region) each participant is focusing on, and 6) recognising the identity of each participant.
4. *Processing and analysis of derived data*: To research and develop techniques for segmentation, structuring, information retrieval, and summarization of meetings.
5. *Multimedia presentation*: To develop flexible frameworks to access and present streams of multimodal data and metadata.

Progress in these AMI research themes requires a large data set on which empirical observations may be made, and on which technologies may be developed. For example, automatic speech recognition systems require many hours of speech on which statistical models may be trained. A key effort early in the AMI project has thus been the production of the AMI Meeting Corpus, consisting of 100 hours of meetings data suitable for

research. The remainder of this article details the design, collection, annotation and distribution of the corpus.

3 Corpus Design

Any study of naturally-occurring behaviour such as meetings immediately encounters a well-known methodological problem: if one simply observes behaviour "in the wild", one's results will be difficult to generalize, since not enough will be known about what is causing the individual(s) to produce the behaviour. [1] identifies seven factors that affect how work groups behave, ranging from the means they have at their disposal, to aspects of organisational culture and pressures coming from the external environment. The type of task the group is trying to perform, and the particular roles and skills the group members bring to it, play a large part in determining what the group does; for instance, if the group members have different roles or skills that bear on the task in different ways, that can increase the importance of some contributions, and be a deciding factor in whether the group actually needs to communicate or can simply leave one person to do the work. Variations to any of the above-mentioned factors will cause the data to change in character, but using observational techniques, it is difficult to get enough of a group history to tease out these effects.

One response to this dilemma is not to make completely natural observations, but to elicit data in a manner which controls as many of these factors as possible. Experimental control allows the researcher to find effects with greater clarity and confidence than in observational work. This approach, well-established in psychology and familiar from some existing corpora (e.g. [2]), comes with its own danger: results obtained in the laboratory will not necessarily generalise to the outside world, since people may simply behave differently when performing an artificial task.

Our response to this methodological difficulty is to collect our data-set in two parts: elicited scenario-driven data, and natural data. The first part (approximately 65 hours) consists of material elicited using a design task in which an effort is made to control the factors from [1]. Since this is the larger part of the data-set, the details of how it was elicited are important, and so we describe it below. The second part (approximately 35 hours) contains naturally occurring meetings in a variety of types, the purpose of which is to help us validate our findings from the elicitation and test their generality. We note that, in fact, a third type of data is also collected, consisting of data elicited using less controlled scenarios than the one which we will describe in this article.

3.1 Meeting elicitation scenario

In our meeting elicitation scenario [3], the participants play the roles of employees in an electronics company that decides to develop a new type of television remote control because the ones found in the market are not user friendly, as well as being unattractive and old-fashioned. The participants are told they are joining a design team whose task, over a day of individual work and group meetings, is to develop a prototype of the new remote control. We chose design teams for this study for several reasons. First, they have functional meetings with clear goals, making it easier to measure effectiveness and efficiency. Second, design is highly relevant for society, since it is a common task in many industrial companies and has clear economic value. Finally, for all teams, meetings are not isolated events but just one part of the overall work cycle, but in design teams, the participants rely more heavily on information from previous meetings than in other types of teams, and so they produce richer possibilities for the browsing technology we are developing.

3.2 Participants and Roles

Within this context, each participant in the elicitation is given a different role to play. The project manager (PM) coordinates the project and has overall responsibility. His job is to guarantee that the project is carried out within time and budget limits. He runs the meetings, produces and distributes minutes, and produces a report at the end of the trial. The marketing expert (ME) is responsible for determining user requirements, watching market trends, and evaluating the prototype. The user interface designer (UI) is responsible for the technical functions the remote control provides and the user interface. Finally, the industrial designer (ID) is responsible for designing how the remote control works including the componentry. The user interface designer and industrial designer jointly have responsibility for the look-and-feel of the design.

For this elicitation, we use participants who are neither professionally trained for design work nor experienced in their role. It is well-known that expert designers behave differently from novices. However, using professional designers for our collection would present both economic and logistical difficulties. Moreover, since participants will be affected by their past experience, all those playing the same role should have the same starting point if we are to produce replicable behaviour. To enable the participants to carry out their work while lacking knowledge and experience, they are given training for their roles at the beginning of the task, and are each assigned a (simulated) personal coach who gives sufficient hints by e-mail on how to do their job. Our past experience with elicitations for similar non-trivial team tasks, such as for crisis management teams, suggests that this approach will yield results that generalise well to real groups. We intend to validate the approach for this data collection both by the comparisons to other data already described and by having parts of the data assessed by design professionals.

3.3 The structure of the elicited data

[4] distinguishes four phases in the design process:

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

We use these phases to structure our elicitation, with one meeting per design phase. In real groups, meetings occur in a cycle where each meeting is typically followed by production and distribution of minutes, the execution of actions that have been agreed on, and preparation of the next meeting. Our groups are the same, except that for

practical reasons, each design project was carried out in one day rather than over the usual more extended period, and we included questionnaires that will allow us to measure process and outcomes throughout the day. In future data collections we intend to collect further data in which the groups have access to meeting browsing technology, and these measures will allow us to evaluate how the technology affects what they do and their overall effectiveness and efficiency. An overview of the group activities and the measurements used is presented in fig. 1.

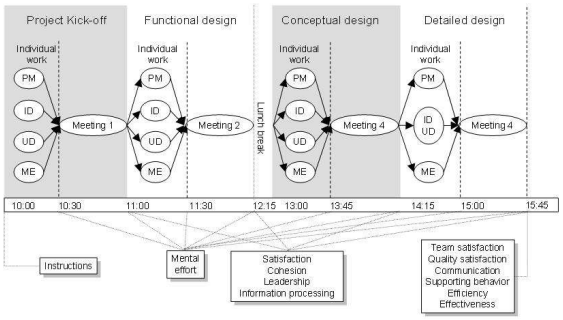

Figure 1: The meeting paradigm: time schedule with activities of participants on top and the variables measured below. PM: Project Manager; ID: industrial designer; UI: user interface designer; ME: marketing expert.

3.4 The working environment

Our collection simulates an office environment in which the participants share a meeting room and have their own private offices and laptops that allow them to send e-mail to each other, which we collect; a web browser with access to a simulated web containing pages useful for the task; and PowerPoint for information presentation. During the trials, individual participants receive simulated e-mail from other individuals in the wider organization, such as the account manager or their head of department, that are intended to affect the course of the task. These emails are the same for every group.

4 Data collection

The data is being captured in three different instrumented meeting rooms that have been built at different project sites. The rooms are broadly similar but differ in shape and construction and therefore in their acoustic properties. In addition, some recording details vary, such as microphone and camera placement and the presence of extra instrumentation. All signals are synchronized by generating a central timecode which is used to replace the timecodes produced locally on each recording device; this ensures, for instance, that videos acquire frames at exactly the same time and that we can find those times on the audio. An example layout, taken from the IDIAP room, is shown in figure 2.

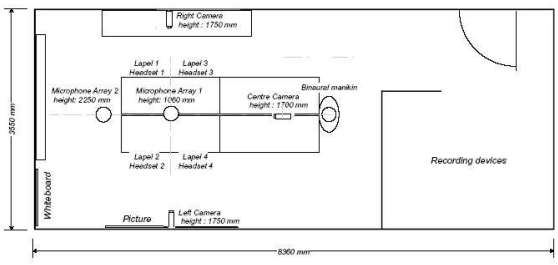

Figure 2: Overhead schematic view of the IDIAP Instrumented Meeting Room.

4.1 Audio

The rooms are set up to record both close-talking and far-field audio. All microphone channels go through separate pre-amplification and analogue to digital conversion before being captured on a PC using Cakewalk Sonar recording software. For close-talking audio, we use omni-directional lapel microphones and headset condenser microphones. Both of these are radio-based so that the participants can move freely. For far-field audio, we use arrays of four or eight miniature omni-directional electret microphones. The individual microphones in the arrays are equivalent to the lapel microphones, but wired. All of the rooms have a circular array mounted on the table in the middle of the participants, plus one other array that is mounted on either the table or the ceiling and is circular in two of the rooms and linear in the other. One room also contains a binaural manikin providing two further audio channels.

4.2 Video

The rooms include capture of both videos that show individuals in detail and ones that show what happens in the room more generally. There is one close-up camera for each of four participants, plus for each room, either two or three room view cameras. The room view cameras can be either mounted to capture the entire room, with locations in corners or on the ceiling, or to capture one side of the meeting table. All cameras are static, with the close-up cameras trained on the participants' usual seating positions. In two of the rooms, output was recorded on Mini-DV tape and then transferred to computer, but in the other output was recorded directly. Figure 3 shows sample output from cameras in the Edinburgh room.

Figure 3: Sample camera views in Edinburgh room

4.3 Auxiliary data sources

In addition to audio and video capture, the rooms are instrumented to allow capture of what is presented during meetings, both any slides projected using a beamer and what is written on an electronic whiteboard. Beamer output is recorded as a time-stamped series of static images, and whiteboard activity as time-stamped x-y co-ordinates of the pen during pen strokes. In addition, individual note-taking uses Logitech I/O digital pens, where the output is similar to what the whiteboard produces. The latter is the one exception for our general approach to synchronization; the recording uses timecodes produced locally on the pen, requiring us to synchronize with the central timecode after the fact as best we can. We intend to subject all of these data sources to further processing in order to extract a more meaningful, character-based data representation automatically [5, 6].

5 Annotation

The data set is being annotated for a range of properties:

- *Speech transcription*, including speaker turn boundaries and word timings,
- *Named entities*, focusing on references to people, artefacts, times, and numbers;

- *Dialogue acts*, using an act typology tailored for group decision-making and including some limited types of relations between acts;
- *Topic segmentation* that allows a shallow hierarchical decomposition into subtopics and includes labels describing the topic of the segment;
- A segmentation of the meetings by the current *group activity* in terms of what they are doing to meet the task in which they are engaged;
- *Extractive summaries* that show which dialogue acts support material in either the project manager's report summarizing the remote control scenario meetings or in third party textual summaries;
- *Emotion* in the style of FeelTrace [11] rated against different dimensions to reflect the range that occurs in the meeting;
- *Head and hand gestures*, in the case of hands focusing on those used for deixis;
- *Location of the individual* in the room and *posture* whilst seated;
- *Location of participant faces and hands* within video frames; and
- *Focus of attention*, i.e. at which other people or artefacts the participants are looking.

For each of these annotations, reliability, or how well different annotators agree on how to apply the schemes, is being assessed. Creating annotations that can be used together for such a wide range of phenomena requires careful thought about data formats, especially since the annotations combine temporal properties with quite complex structural ones, such as trees and referential links, and since they may contain alternate readings for the same phenomenon created by different coders. We use the NITE XML Toolkit for this purpose [12]. Many of the annotations are being created natively in NXT's data storage format using GUIs based on NXT libraries — figure 4 shows one such tool — and others require up-translation, which in most cases is simple to perform. One advantage for our choice of storage format is that it makes the data amenable to integrated analysis using an existing query language.

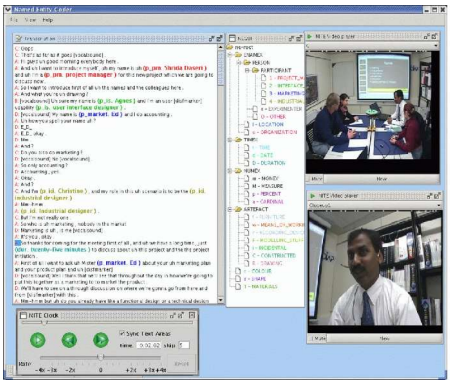

Figure 4: Example annotation GUI in NXT

6 Distribution

Although at the time of submission, the data set has not yet been released, it will become publicly accessible via http://mmm.idiap.ch. The existing Media File Server found there allows users to browse available recorded sessions, download and upload data in a variety of formats, and play media (through streaming servers and players), as well as providing web hosting and streaming servers for the Ferret meeting browser [13].

Supported by *EU 6th FWP IST Integrated Project AMI (FP6-506811, publication AMI-108)*

References:

1. McGrath, J.E., Hollingshead, A.: Interacting with Technology: Ideas, Evidence, Issues and an Agenda. Sage Publications, Thousand Oaks (1994)
2. Anderson, A.H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., Weinert, R.: The HCRC Map Task Corpus. Language and Speech 34 (1991) 351–366
3. Post, W.M., Cremers, A.H., Henkemans, O.B.: A research environment for meeting behavior. In Nijholt, A., Nishida, T., Fruchter, R., Rosenberg, D., eds.: Social Intelligence Design, University of Twente, Enschede, the Netherlands (2004)
4. Pahl, G., Beitz, W.: Engineering design: a systematic approach. Springer, London (1996)
5. Chen, D., Odobez, J.M., Bourlard, H.: Text detection and recognition in images and video frames. Pattern Recognition 37 (2004) 595–608
6. Liwicki, M., Bunke, H.: Handwriting recognition of whiteboard notes. In Marcelli, A., ed.: 12th Conference of the International Graphonomics Society, Salerno (2005)
7. Lathoud, G., McCowan, I.A., Odobez, J.M.: Unsupervised location-based segmentation of multi-party speech. In: ICASSP-NIST Meeting Recognition Workshop,Montreal (2004) http://www.idiap.ch/publications/lathoud04a.bib.
8. Hain, T., Dines, J., Garau, G., Moore, D., Karafiat, M., Wan, V., Oerdelman, R., Renals, S.: Transcription of conference room meetings: an investigation. In: InterSpeech 2005, Lisbon (to appear)
9. Fitt, S.: Documentation and user guide to UNISYN lexicon and post-lexical rules. Technical report, Centre for Speech Technology Research, University of Edinburgh (2000)
10. Greenberg, S.: Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation. In: ESCA Workshop on modelling pronunciation variation for automatic speech recognition, Kerkrade, Netherlands (1998) 47–56
11. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schroder, M.: 'FEELTRACE': An instrument for recording perceived emotion in real time. In Douglas-Cowie, E., Cowie, R., Schroder, M., eds.: ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research, Belfast (2000) 19–24
12. Carletta, J., Evert, S., Heid, U., Kilgour, J., Reidsma, D., Robertson, J.: The NITE XML Toolkit. (submitted)
13. Wellner, P., Flynn, M., Guillemot, M.: Browsing recorded meetings with Ferret. In Bengio, S., Bourlard, H., eds.: Machine Learning for Multimodal Interaction: First International Workshop, MLMI 2004, Martigny, Switzerland, June 21-23, 2004, Revised Selected Papers. Lecture Notes in Computer Science 3361. Springer- Verlag, Berlin (2005) 12–21