A New Generation of Meeting Browsers

Dr Jean-Albert Ferrez and Dr Iain McCowan, IDIAP Research Institute, Martigny
(contact: jean-albert.ferrez@idiap.ch)

Jean-Albert Ferrez and Iain McCowan

We all have too many meetings, often with nothing to show for them afterwards except sketchy minutes or hastily written notes. Basic technologies such as hands-free speakerphones and video-conferencing equipment, beamers and Powerpoint presentations, audio and video recorders, electronic whiteboards etc. all exist to make our lives easier when we attend meetings. Researchers are working on new approaches to the capture, archiving, summarization and delayed/remote viewing of meetings that will significantly improve access to the information generated and exchanged – and too often forgotten – during those meetings.

This research revolves around instrumented meeting rooms which allow the collection, annotation, structuring, and browsing of multimodal meeting records. For each meeting, audio, video, slides and textual information (notes, whiteboard text, etc.) are recorded and time-synchronized. Relevant information is extracted from these raw multimodal signals using state-of-the-art processing technologies. The resulting multimedia and information streams are then available to be structured, browsed and searched within an easily accessible archive.

The AMI (Augmented Multi-party Interaction)[6] European project, coordinated by IDIAP, is particularly concerned with the application of multimodal processing technologies to develop meeting browsers and remote meeting assistants. A meeting browser is a system that enables a user to navigate an archive of meetings, viewing and accessing the entire multimodal content based on automatic annotation, structuring and indexing of the information streams. For example, navigation may be enabled using automatic annotations such as speech transcription or identification of participants. A natural extension of such a meeting browser is the concept of a remote meeting assistant which performs such operations in real time during a meeting and enables remote participants to have a much richer interaction with the meeting.

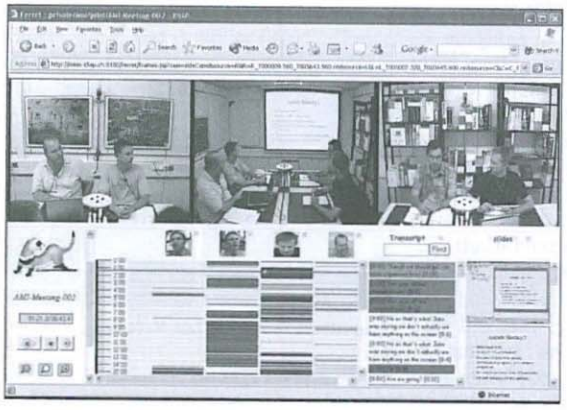
To develop these applications, the AMI project is pushing back the frontiers in several areas including the modelling of group dynamics, the processing and recognition of audio and visual signals, the creation of models that combine multiple modalities, the abstraction of content from multi-party meetings, and human-computer interaction. These R&D themes are underpinned by the ongoing capture of user requirements, the development of a common infrastructure, and the evaluation 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 supersede any single domain of application. Each of the above technologies has a broad application, for example in security, surveillance, home care monitoring and in more natural human-computer interfaces.

---

[6] Cf. www.amiproject.org.

Definition and Analysis of Meeting Scenarios

To design appropriate technologies for meetings it is first necessary to understand the type of participants, the nature of their interactions and the means by which they communicate. In order to provide a structured framework for research, rather than studying a variety of unconstrained meetings our initial work is focusing on investigating scenario-based meetings. These are motivated by a scenario which is given to participants beforehand to describe aspects relating to the meeting as a whole, such as its purpose, theme, context, and expected duration. Participants in these meetings act naturally as themselves, but assume an artificial role (e.g. project manager). Use of such scenarios follows a standard methodology taken from social psychology, allowing us to study groups that behave as much as possible like real groups of the desired type whilst still being able to control experiment conditions.



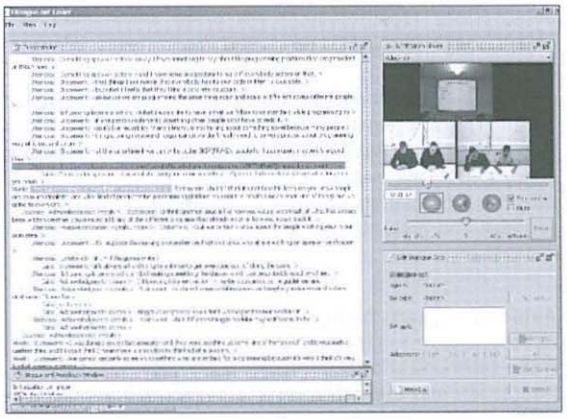
Prototype AMI meeting browser

Within the framework of the ongoing projects, a scenario has been defined based on a series of meetings in a design project. Studying a series of meetings within the context of a project allows us to improve our understanding of the processes that occur not only during meetings, but also between them – meetings take place within a particular environment and are generally part of an ongoing work cycle. Design project meetings have the advantage of a strong natural structure, generally progressing through the phases of brainstorming, negotiation and decision-making, which makes them suitable for automated techniques for structuring and summarizing information. In addition, design project meetings work towards a quantifiable outcome, allowing for clearer evaluation of the impact of social and organizational factors and also of any assisting technologies that are being employed. To complement data collected in this scenario-driven manner, and to confirm the applicability of research results, other real meetings are also being collected and analyzed.

Automatic Extraction of Information from Audio-Visual Data

The focus of the AMI project is to add value to meetings by automatically documenting and presenting their information content. To achieve this, meetings are recorded and observed using microphones and video cameras, and then state-of-the-art technologies are applied to extract the raw information content from these audio-visual signals. AMI technologies processing these signals may be grouped according to the six core types of information that they extract: (1) recognizing what participants say (automatic speech recognition, see below), (2) recognizing what participants do (automatic action and gesture recognition), (3) recognizing

where each participant is each time (localization and tracking), (4) recognizing how participants act from the point of view of their emotional state (emotion recognition, see below), (5) tracking which person, object or region each participant is focusing on (focus-of-attention recognition), and (6) recognizing the identity of each participant (person recognition).



Tool developed to annotate meeting dialogue acts

So far, significant progress has been made towards solving each of the above problems using state-of-the-art signal processing and machine learning technologies. Signal processing takes raw audio-visual data and reformats it in some way, for example extracting voice information from background noise, while machine learning techniques automatically recognize patterns present in large quantities of data, for example learning what a word sounds like by hearing it said many times. In particular work has focused on applying techniques to data recorded in real-world conditions and on using both audio and visual information whenever this is relevant and complementary (for example using both vocal and facial information to identify someone).

Output of an audio-visual, multi-view, multi-person tracker

Automatic Structuring and Summarization of Information

Meetings contain much raw information in the forms described above, such as spoken or written words, gestures, actions and emotions. To allow efficient access to relevant parts of this information, it is necessary to provide some higher-level structure or to distil the core information in the form of a summary. Various ways of structuring meeting information are being developed, for example according to dialogue structure, different topic categories, or phases of

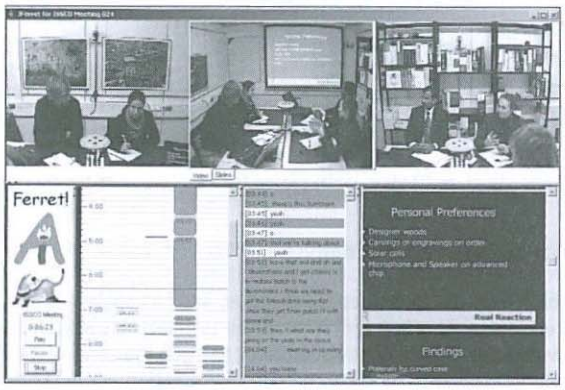
the meeting, such as discussions or presentations. Different approaches to summarization are also being studied, including extractive summarization (where only the key informative segments of the meeting are identified) and abstractive summarization (where a coherent high-level text is generated to describe the most important information from the meeting).

Multimedia Retrieval and Presentation

The automatic information extraction technologies described above aim to supplement the raw multimedia meeting recordings with various types of metadata – words, identities, actions, summaries etc. A further direction for research is to develop technologies that allow this multimedia data and metadata to be presented to an end-user in ways that allow him to retrieve relevant information. Researchers at IDIAP are focusing on two target applications for these retrieval and presentation technologies, namely a meeting browser and a remote meeting assistant.

Automatic Speech Recognition System

While communication in meetings is multimodal in nature, speech is the predominant mode of interaction and the richest source of information. Compared to other domains of application such as dictation or broadcast news transcription, meetings pose a number of challenges for Automatic Speech Recognition (ASR) systems. Speech in meetings is conversational in nature, and so does not follow standard grammatical constraints; we are addressing this by modelling the informal grammar of conversational speech to include for example corrections and repetitions. There are also difficulties caused by having several people speaking in the same room – turns at speaking can be difficult to segment, and people often talk over the top of one another. By processing multiple microphones jointly, researchers are developing techniques to segment and separate concurrent speech. A further challenge is the occurrence of non-native speech – while English is often used in international business meetings, it is spoken with differing accents and with a variable degree of fluency; researchers are investigating pronunciation models that take account of such variations. Finally, while current ASR systems rely on headset microphones, there is a need to move towards less constraining hands-free microphones; we are investigating the use of table-top microphone arrays which automatically track and focus sound acquisition on a particular speaker. As these new techniques emerge, they are gradually integrated into a state-of-the-art large vocabulary conversational ASR system.



JFerret multimedia browser toolkit

The JFerret Multimedia Browser Toolkit

JFerret is a new multimedia browser developed at IDIAP which gives researchers a common platform for developing integrated presentation

demonstrators. The browser is extremely flexible, allowing almost any user interface to be composed using a combination of plug-in modules. An XML (Extensible Markup Language) configuration file specifies which plug-in components to use, how to arrange them visually and how they will communicate with each other. The picture here shows a sample browser implemented with JFerret – it uses 29 plug-ins, including three videos and an audio player, but is configured in less than one page of XML.

JFerret comes with a library of pre-defined plug-ins for presentation of video, audio, slides, annotation time-lines, controls and more. This base set of presentation components can easily be extended by writing new plug-ins. JFerret is written in Java and provides a simple plug-in programming interface and an elegant communication mechanism between plug-ins. Java also allows the application to run cross-platform, either as an Applet or as a stand-alone application.

The National Centre of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2)

IM2 is aimed at the advancement of research and the development of prototypes in the field of man-machine interaction. The NCCR is particularly concerned with technologies coordinating natural input modes (such as speech, image, pen, touch, hand gestures, head and/or body movements, and even physiological sensors) with multimedia system outputs such as speech, sounds, images and animation.

The field of multimodal interaction covers a wide range of activities and applications including the recognition and interpretation of spoken, written and gestured languages, computer vision, and the automatic indexation and management of multimedia documents. Other important related themes are information content protection, data access control, and the structuring, retrieval and presentation of multimedia information.

Multimodal interfaces represent a new, highly strategic direction for information technologies of the future. Thanks to such interfaces, man-machine interactions will become simpler and in consequence, more productive. In the near future, multimedia systems equipped with such interfaces will be flexible enough to accommodate a wide variety of users, tasks and environments for which current interaction modalities (such as keyboard, mouse and screen) are inadequate. In the first instance, ideal interfaces would be capable of manipulating more complex and more realistic data, including the combination of different forms of data such as audio and video.

IM2 is led by the IDIAP Research Institute in Martigny, and combines many partners from a number of university institutions (EPFL, University of Geneva, University of Fribourg, University of Bern, ETHZ), along with HES-SO (Sion, Sierre, Fribourg, Lausanne) and a range of commercial companies. The NCCR also has numerous international contacts including an agreement for the exchange of young researchers with ICSI in Berkeley, California.

Emotion in Meetings

A full understanding of human communication cannot be achieved without some indication of its emotional content. When a decision was taken, were people disappointed or pleased? Were they relaxed or nervous when a particular question was asked? IDIAP researchers are investigating the definition of emotional content in meetings, as well as automatic techniques for its classification. To begin with, a study was conducted to determine the types of emotion commonly displayed in meetings: findings included the emotional adverbs listed above, along with others such as angry or relaxed. It is, however, difficult to accurately categorize natural emotions, and so researchers are instead adopting a dimensional approach to labelling emotions. Participants are continually rated along two dimensions: one axis indicates whether the emotion is negative (e.g. anger) or positive (e.g. pleased), while the other shows whether the person is exhibiting the emotion in a passive (e.g. bored) or an active manner (e.g. joking). Algorithms have been developed to categorize or rate emotional

content from a person's audio cues such as speech pitch or volume and visual cues such as facial expression, and linguistic information has also been used to improve the accuracy of automatic emotion recognition. Approaches have also been developed to recognize meeting hot-spots, the periods when participants exhibit a high degree of involvement or interest.

From Lab to Market

Some of the components discussed above are in a very advanced state and already being exploited in usable prototypes. Other components still require much further work and a few theoretical breakthroughs before they can be transferred from the labs to the end-user. Products available today already implement some of the ideas and concepts developed above, such as recording and web-based playback of conversations. Key industrial players in the areas of information technology and telecommunications are keeping a close watch on this research and have expressed strong interest in incorporating the latest results in new products. However, many classic meetings have yet to take place before we can rely on a fully automatic, artificial but intelligent meeting assistant that will completely eliminate the need for physical presence and personal note-taking.