Augmented **M**ulti-party **I**nteraction
http://www.amiproject.org

Augmented **M**ulti-party **I**nteraction with **D**istance **A**ccess
http://www.amidaproject.org

AMI overview and prospects for future research

(January 2006)

Information Society
Technologies

Sixth Framework Programme

1 Introduction

As networks and computers become more pervasive, groups are increasingly using technology to assist communication and collaboration. Computer assistance should have the potential to make groups work more effectively, particularly in the increasing number of cases where group members are not all in the same place at the same time. Some technologies to support remote group working, such as teleconferencing and videoconferencing, are well established. Hundreds of millions of Euros have been invested in their development. They are in common use, and they are improving all the time. However, industrial experience suggests that although these technologies may be better for remote groups than relying on asynchronous options like email, communication using them is greatly impaired compared with face-to-face discussion.

This is for two reasons. Firstly, even under good conditions, it is hard to convey anything beyond the bare words: tone and nuance disappear in degraded audio, and only very high quality, properly synchronized video could possibly convey the many and subtle visual cues about meaning. Secondly, taking turns, or even knowing who said what, can be difficult, again because the multimodal cues we most rely on in face-to-face discussion are missing when using these technologies. Even face-to-face groups find it hard to have coherent, efficient discussions and transfer the results into some kind of institutional memory. For remote groups, the barriers are even higher, but the potential benefits to be found in addressing this problem are much greater.

Technology can also be used to support groups working on large, complex, multimodal datasets. The sophisticated virtual and immersive environments now deployed in various settings are are able to provide high quality visualization of data, together with some facilities for group communication (typically commercially available videoconferencing technology).

What groups need is not just the basic infrastructure that allows them to hold remote meetings, but thoughtfully designed technologies that recognize and overcome the communication and collaboration difficulties that they face. FP6 integrated projects such as AMI and CHIL are addressing these issues, based on the use of "ambient" human-computer interfaces that are able to recognize and interpret different modalities of human communication, at both the individual and group level, and to maintain and track context in communicative scenes and dynamically changing environments. These projects are making significant advances both in terms of foundational technologies (such as speech recognition, visual scene analysis, multimodal fusion, and content abstraction) and in terms of demonstration systems.

However, we are still far from the integrated collaborative environments that promise dramatically improved productivity through the provision of better interfaces, better collaboration, and better data, enabling better decisions to be made faster. To achieve these goals requires a major research effort in the analysis and understanding of *communication scenes*, and the development of *mixed reality collaborative environments* built on ambient interfaces.

2 Augmenting meetings

Started in January 2004, the European AMI (Augmented Multi-party Interaction) Integrated Project has been building systems to enhance the way meetings are run and documented.

AMI research revolves around instrumented meeting rooms which enable the collection, annotation, structuring, and browsing of multimodal meeting recordings. 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 queried within an easily accessible archive.

AMI is particularly concerned with the application of multimodal processing technologies to develop meeting browsers and remote meeting assistants. A meeting browser (as illustrated below) is a system that enables a user to navigate (interaction) an archive of meetings, viewing (visualization) and accessing the full multimodal content, based on automatic annotation, structuring and indexing of the information streams. While initial research and evaluation efforts were performed on“scripted meetings” (simulation), solutions are now being developed for real life (face-to-face) meetings. Soon, meeting rooms will also be connected to allow for remote collaboration, involving additional tools, such as shared workspaces/“whiteboards” (mixed reality).

Computational models to automatically recognize meeting features are built using machine learning algorithms and large sets of training data. The performance of these models is then evaluated and improved through systematic simulation using recorded test data to determine how well these systems will perform on real meetings.

One possible “vision”, as currently developed in AMI, is of a “remote meeting assistant”, illustrated in Figure 1, that can understand what is happening well enough to tell someone when topics of interest come up, brief them on what has happened so far, and help them cope with poor connections.

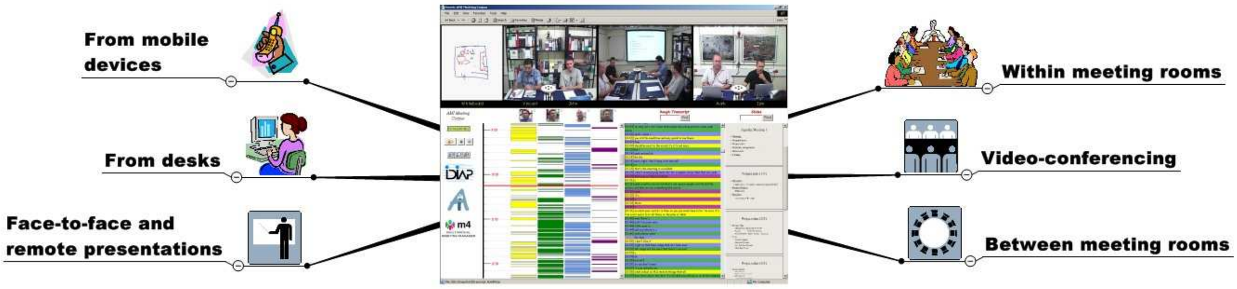

Figure 1: *AMI remote meeting assistant: The focus for AMI is natural, unconstrained and multimodal human communication, from anywhere to anywhere, and using whatever interaction devices and network bandwidth is available. This includes face-to-face and remote meetings, whether they are one-to-many (as in remote presentations), many-to-many (e.g., between meeting rooms), or something in between. Shown at the center of the figure is the AMI JFerret browser, which currently underpins our support for archived face-to-face meetings. The top panel plays videos and whiteboard strokes as they are made. Below is a vertical timeline, with coloured bars showing when participants talked, plus an ASR transcript and any slides, which can be searched for keywords. Clicking on any item jumps playback to the corresponding moment in the meeting.*

3 Understanding communication scenes

Human communication is complex and is factored across several modalities. To address the problem requires RTD effort in several traditionally separate disciplines including unconstrained speech recognition, visual scene analysis, modelling individuals and groups through the joint processing of multiple information channels, and structuring, indexing and summarizing these multimodal communication scenes. Projects such as AMI and CHIL have made significant progress in these basic areas, and new projects will further push the state-of-the-art (eg in terms of realtime processing). The scientific outputs of these projects (technology components, large annotated databases, evaluation protocols) will form a platform for future research in the area (which is one in which Europe has a lead).

These foundational ways of processing communication scenes will underpin the development of technologies for effective collaboration. Following studies by Chuck House of Intel, we factor the problem into three strands:

Archives: Technologies to create archives are centred on the multimodal recordings processed, structured and indexed using speech recognition, computer vision, linguistic and discourse modelling, content abstraction, and interaction modelling. This is a focus of projects such as AMI and CHIL.

Context: Effective collaboration requires a lot of background context, such as automatic access to relevant archives, personalized information presentations, realtime access to the state of the group

Presence: The biggest problem with videoconferencing and other existing remote meeting technologies is that you are not *there*. Presence can simply involve broadcasting your present state (as rich as the technology allows) but it can build on attentional cues, group dynamics, effective interaction with shared data, etc.

4 Mixed reality collaborative environments

Currently we are seeing ever-increasing network bandwidth (particularly to homes and to mobile users), the introduction of better display technologies now targetted at consumers, and the deployment fast, parallel processors such as the IBM/Sony/Toshiba Cell Processor, which specifically focuses on media processing. Immersive environments (eg those developed by SGI) are now a high end technology, used in relatively few facilities, but they have the potential to become mainstream consumer technology within a few years.

We should exploit these trends when considering how to support human-human interaction (and collaborative interaction with data). Most of our current interfaces to data and to collaboration are not sufficiently rich to provide the context and presence that we require (and are starting to be able to provide).

A mixed reality collaborative environment, based on immersive environments has the potential to transform the way we meet (face-to-face and remote), and the way we interact with data.

Context: Immersive 3D environments have the capability to make large amounts of data available, by creating peripheral context—information doesn't disappear, it just goes to the background. And as new data emerges, or as the conversation evolves, so the ambient interface is able to project newly important data to the foreground. Access to archives becomes transparent, automatic and natural.

Presence: Constructing a collaborative workspace for both people and data, supports the feeling of presence in a remote collaboration. The collaborative space can be structured according the focus of attention of participants, and the interactive dynamics.