1 Multimodal signal processing for meetings: an introduction

Andrei Popescu-Belis and Jean Carletta

This book is an introduction to multimodal signal processing. In it, we use the goal of building applications that can understand meetings as a way to focus and motivate the processing we describe. Multimodal signal processing takes the outputs of capture devices running at the same time – primarily cameras and microphones, but also electronic whiteboards and pens – and automatically analyzes them to make sense of what is happening in the space being recorded. For instance, these analyses might indicate who spoke, what was said, whether there was an active discussion, and who was dominant in it. These analyses require the capture of multimodal data using a range of signals, followed by a low-level automatic annotation of them, gradually layering up annotation until information that relates to user requirements is extracted.

Multimodal signal processing can be done in real time, that is, fast enough to build applications that influence the group while they are together, or offline – not always but often at higher quality – for later review of what went on. It can also be done for groups that are all together in one space, typically an instrumented meeting room, or for groups that are in different spaces but use technology such as videoconferencing to communicate. The book thus introduces automatic approaches to capturing, processing, and ultimately understanding human interaction in meetings, and describes the state of the art for all technologies involved.

Multimodal signal processing raises the possibility of a wide range of applications that help groups improve their interactions and hence their effectiveness between or during meetings. However, developing applications has required improvements in the technological state of the art in many arenas.

The first arena comprises core technologies like audio and visual processing and recognition that tell us basic facts such as who was present and what words were said. On top of this information comes processing that begins to make sense of a meeting in human terms. Part of this is simply combining different sources of information into a record of who said what, when, and to whom, but it is often also useful, for instance, to apply models of group dynamics from the behavioral and social sciences in order to reveal how a group interacts, or to abstract and summarize the meeting content overall. Finding ways to integrate the varying analyses required for a particular meeting support application has been a major new challenge.

Multimodal Signal Processing: Human Interactions in Meetings, ed. Steve Renals, Hervé Bourlard, Jean Carletta, and Andrei Popescu-Belis. Published by Cambridge University Press.

Finally, moving from components that model and analyze multimodal human-to-human communication scenes to real-world applications has required careful user requirements capture, as well as interface and systems design. Even deciding how to evaluate such systems breaks new ground, whether it is done intrinsically (that is, in terms of the accuracy of the information the system presents) or from a user-centric point of view.

1.1 Why meetings?

The research described in this book could be applied to just about any setting where humans interact face-to-face in groups. However, it is impossible to design reasonable end-user applications without focusing on a specific kind of human interaction. Meetings provide a good focus for several reasons.

First, they are ubiquitous. Meetings pervade nearly every aspect of our communal lives, whether it is in work, in the running of community groups, or simply in arranging our private affairs. Meetings may not be the only way in which humans interact, but they are a frequent and understandable one, with obvious practical relevance.

Second, what happens in meetings (or, as often, what does not) is actually important. For many people, meetings are the milestones by which they pace their work. In truly collaborative decision-making, the meeting is where a group's goals and work take shape. Even in groups where the real decision-making takes place behind the scenes, in the absence of written documents the meeting itself is where a group's joint intention is most fully and most clearly expressed. Being able to understand what happens in meetings is bound to be useful, whether the goal is to reveal the content of the meeting or simply to identify where a group's process could be improved.

Third, because of changes in modern society, meetings present an obvious opportunity. Many organizations operate globally. There are few jobs for life. In the face of staff churn and business fragmentation, it is increasingly difficult for organizations simply to keep and access the institutional memory they need in order to make good decisions. Adequately documenting everything in writing is expensive, if not impossible. This makes it economically important to get better control of the information locked in meetings, starting from adequate options to record, analyze, and access some of the media related to them.

Finally, a great many meetings take place in settings where there is already, or is developing, a sense that the benefits of recording outweigh privacy considerations. Many organizations already record and archive at least their key meetings routinely, even without decent tools for sifting later through what they have stored. This is not just a matter of the technology for recording being cheap enough (although of course this is a factor), but of the organizations hoping to function better thanks to the recordings. This in itself brings benefits for an organization's members, but there can be more personal benefits too. Meetings may be ubiquitous, but we cannot always be at all of the ones that affect us. Being able to glean their content efficiently is likely to help.

1.2 The need for meeting support technology

Like 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. 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-connected 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.

Meetings play a crucial role in the generation of ideas, documents, relationships, and actions within an organization. Traditionally, depending on the type of meeting, either everyone will take whatever style of notes they please, or one person will create official written minutes of the meeting. Whatever the form of written record, it will be subjective and incomplete. Even with the best minutes, business questions often appear later, which can only possibly be resolved by going back to what actually happened. The technology now exists to capture the entire meeting process, keeping the text and graphics generated during a meeting together with the audio and video signals.

If only people could use the multimedia recordings of meetings to find out or remember what they need to know about the outcome of a meeting, then using these recordings would become an attractive adjunct (or even, alternative) to note taking. This can only happen once it is possible to recognize, structure, index, and summarize meeting recordings automatically so that they can be searched efficiently. One of the long-term goals of meeting support technology is to make it possible to capture and analyze what a group of people is doing together in a room-sized space using portable equipment, and to put together a wide range of applications supporting the group, using configurable componentry or web services for tasks like recognizing the speech, summarizing, and analyzing the group's interaction. This will enable companies to make use of archives of meetings, for instance, for audit purposes or to promote better cohesion in globalized businesses. Different configurations of the same underlying components will also help people who work away from the office to participate more fully in meetings. These possibilities indicate that we are at the point of a big technological breakthrough.

1.3 A brief history of research projects on meetings

The ideas presented in this book stem for a large part, though not exclusively, from the contributions made by the members of the AMI Consortium. This network of research and development teams was formed in the year 2003 building upon previous collaborations. However, several other large initiatives focused as well on multimodal signal processing and its application to meeting analysis and access, and were either precursors or contemporaries of AMI.

1.3.1 Approaches to meeting and lecture analysis

The understanding of human communication has long been a theoretical goal of artificial intelligence, but started having also practical value for information access through the 1990s, as more and more audio-visual recordings were available in digital formats. During the 1990s, separate advances in the audio and video analysis of recordings led to the first implemented systems for interaction capture, analysis, and retrieval. The early Filochat system (Whittaker *et al.*, 1994b) took advantage of handwritten notes to provide access to recordings of conversations, while BBN's Rough'n'Ready system (Kubala *et al.*, 1999) enhanced audio recordings with structured information from speech transcription supplemented with speaker and topic identification. Video indexing of conferences was also considered in early work by Kazman *et al.* (1996). Multichannel audio recording and transcription of business or research meetings was applied on a considerably larger scale in the Meeting Recorder project at ICSI, Berkeley (Morgan *et al.*, 2001, 2003), which produced a landmark corpus that was reused in many subsequent projects.

Around the year 2000, it became apparent that technologies for meeting support needed to address a significant subset of the modalities actually used for human communication, not just one. This in turn required appropriate capture devices, which needed to be placed in instrumented meeting rooms, due to constraints on their position, size, and connection to recording devices, as exemplified by the MIT Intelligent Room with its multiple sensors (Coen, 1999). The technology seemed mature enough, however, for corporate research centers to engage in the design of such rooms and accompanying software, with potential end-user applications seeming not far from reach.

For instance, Classroom 2000 (Abowd, 1999) was an instrumented classroom intended to capture and render all aspects of the teaching activities that constitute a lecture. The Microsoft Distributed Meetings system (Cutler *et al.*, 2002) supported live broadcast of audio and video meeting data, along with recording and subsequent browsing. Experiments with lectures in this setting, for example for distance learning, indicated the importance of video editing based on multimodal cues (Rui *et al.*, 2003). Instrumented meeting or conference rooms were also developed by Ricoh Corporation, along with a browser for audio-visual recordings (Lee *et al.*, 2002), and by Fuji Xerox at FXPAL, where the semi-automatic production of meeting minutes, including summaries, was investigated (Chiu *et al.*, 2001).

However, even if companies were eager to turn meeting support technology into products, it became clear that in order to provide intelligent access to multimedia recordings of human interaction a finer-grained level of content analysis and abstraction was required, which could simply not be achieved with the knowledge available around the year 2000. Technology for remote audio-visual conferencing has been embedded into a host of successful products,[1] but without analyzing the conveyed signals and generally with highly limited recording or browsing capabilities.

[1] To name but a few: HP's Halo (now owned by Polycom) or CISCO's WebEx for the corporate market, and Skype, iChat, or Adobe Connect as consumer products.

1.3.2 Research on multimodal human interaction analysis

The need for advanced multimodal signal processing for content abstraction and access has been addressed in the past decade by several consortia doing mainly fundamental research. Only such collaborative undertakings could address the full complexity of human interaction in meetings, which had long been known to psychologists (e.g., Bales, 1950, McGrath, 1984). Moreover, only such consortia appeared to have the means to collect large amounts of data in normalized settings and to provide reference annotations in several modalities, as needed for training powerful machine learning algorithms. The public nature of most of the funding involved in such initiatives ensured the public availability of the data.

Two projects at Carnegie Mellon University (CMU) were among the first to receive public funding to study multimodal capture, indexing, and retrieval, with a focus on meetings. The target of the Informedia project was first the cross-modal analysis of speech, language, and images for digital video libraries (1994–1999), and then the automatic summarization of information across multimedia documents (1999–2003) (Wactlar *et al.*, 1996, 2000). In parallel, CMU's Interactive Systems Laboratory initiated a project on meeting record creation and access (Waibel *et al.*, 2001a). This was directly concerned with recording and browsing meetings based on audio and video information, emphasizing the role of speech transcription and summarization for information access (Burger *et al.*, 2002).

In Europe, the FAME project (Facilitating Agent for Multicultural Exchange, 2002–2005) developed the prototype of a system that made use of multimodal information streams from an instrumented room (Rogina and Schaaf, 2002) to facilitate cross-cultural human–human conversation. A second prototype, the FAME Interactive Space (Metze *et al.*, 2006), provided access to recordings of lectures via a table top interface that accepted voice commands from a user. The M4 European project (MultiModal Meeting Manager, 2002–2005), introduced a framework for the integration of multimodal data streams and for the detection of group actions (McCowan *et al.*, 2003, 2005b), and proposed solutions for multimodal tracking of the focus of attention of meeting participants, multimodal summarization, and multimodal information retrieval. The M4 Consortium achieved a complete system for multimodal recording, structuring, browsing, and querying an archive of meetings.

In Switzerland, the IM2 National Center of Competence in Research is a large long-term initiative (2002–2013) in the field of Interactive Multimodal Information Management. While the range of topics studied within IM2 is quite large, the main application in the first two phases (2002–2009) has focused on multimodal meeting processing and access, often in synergy with the AMI Consortium. The IM2 achievements in multimodal signal processing (see for instance Thiran *et al.*, 2010) are currently being ported, via user-oriented experiments, to various collaborative settings.

Two recent joint projects were to a certain extent parallel to the AMI and AMIDA projects. The CHIL European project (Computers in the Human Interaction Loop, 2004–2007) has explored the use of computers to enhance human communication in smart environments, especially within lectures and post-lecture discussions, following

several innovations from the CMU/ISL and FAME projects mentioned above (Waibel and Stiefelhagen, 2009). The US CALO project (Cognitive Assistant that Learns and Organizes, 2003–2008) has developed, among other things, a meeting assistant focused on advanced analysis of spoken meeting recordings, along with related documents, including emails (Tür *et al.*, 2010). Its major goal was to learn to detect high-level aspects of human interaction which could serve to create summaries based on action items.

It must be noted that projects in multimodal signal processing for meetings appear to belong mainly to three lineages: one descending from CMU/ISL with the FAME and CHIL projects (with emphasis on lectures, video processing and event detection), another one from ICSI MR to CALO (with emphasis on language and semantic analysis), and finally the lineage from M4 and IM2 to AMI and AMIDA (with a wider and balanced approach). Of course, collaborations between these three lineages have ensured that knowledge and data have moved freely from one to another.

1.3.3 The AMI Consortium

The technologies and applications presented in this book are closely connected to the research achievements of the AMI Consortium, a group of institutions that have advanced multimodal signal processing and meeting support technology. The AMI Consortium was constituted around 2003, building on existing European and international expertise, and on previous collaborations. The consortium was funded by the European Union through two successive integrated projects: Augmented Multiparty Interaction (AMI, 2003–2006) and Augmented Multiparty Interaction with Distance Access (AMIDA, 2006–2009). As a result, the consortium was highly active for more than seven years, which represents a particularly long-term multi-disciplinary research effort, surpassed only by certain national initiatives such as the Swiss IM2 NCCR (twelve years). This book presents only a selection of what the AMI Consortium has achieved, but also includes relevant advances made by the wider research community.

The AMI Consortium has included both academic partners (universities and not-for-profit research institutes) and non-academic ones (companies or technology transfer organizations). Although the partnership has varied over the years, the academic partners were the Idiap Research Institute, the University of Edinburgh, the German Research Center for AI (DFKI), the International Computer Science Institute (ICSI, Berkeley), the Netherlands Organization for Applied Scientific Research (TNO), Brno University of Technology, Munich University of Technology, Sheffield University, the University of Twente, and the Australian CSIRO eHealth Research Center. The primary non-academic partners were Philips and Noldus Information Technology. Interested companies who were not project partners were able to interact with the AMI Consortium through the AMI Community of Interest and in focused "mini-project" collaborations. These interactions allowed industry to influence the research and development work based on market needs and to prepare to use AMI technology within existing or future products and services.

1.3.4 Joint evaluation and dissemination activities

In many fields, the existence of a shared task – with standardized data sets and evaluation metrics – has served as a driving force to ensure progress of the technology. Shared tasks offer an accurate comparison of methods at a given time. They also provide training and test data, thus lowering the entry cost for new institutions interested in solving the task. Shared tasks and standardized evaluation began in 1988 for automatic speech recognition, and since then, the approach has spread more widely.

For multimodal signal processing applied to meetings or lectures, two initiatives have promoted shared tasks: the Rich Transcription (RT) evaluations and the Classification of Events Activities and Relationships (CLEAR) ones. In both series, the US National Institute for Standard Technology (NIST) has played a pivotal role in gathering normalized data that was considered by participants to be representative of the addressed research questions. Along with external data from the AMI and CHIL consortia, NIST has also produced original data in its own instrumented meeting rooms, starting from the Smart Spaces Laboratory (Stanford *et al.*, 2003).

The NIST annual RT evaluations started as early as 2001 for broadcast news and telephone conversations, and meetings were targeted starting 2004. Following increasing interest, the most visible results were produced in the 2005–2007 campaigns, the latter one being organized and published jointly with CLEAR (Stiefelhagen *et al.*, 2008); a smaller workshop was further held in 2009. The goal of the RT evaluations was to compare the performance of systems submitted by participants on meetings of varying styles recorded using multiple microphones. The systems were mainly for automatic speech recognition (producing text from speech, including punctuation and capitalization) and for speaker diarization (determining who spoke when). RT differed from other campaigns for speech recognition, such as broadcast news, in its emphasis on multiple, simultaneous speakers and on non-intrusive capture devices, but did not target higher-level information extraction capabilities on meeting signals, such as those developed by AMI or CALO.

The CLEAR evaluations were sponsored by the US VACE program (Video Analysis and Content Extraction) with support from CHIL and an infrastructure provided by NIST. The CLEAR 2006 and 2007 evaluations (Stiefelhagen and Garofolo, 2007, Stiefelhagen *et al.*, 2008) targeted mainly the problems of person and face tracking, head pose estimation, and acoustic event detection using signals from several capture devices (cameras, microphones) in instrumented meeting rooms. Several conditions were tested for each track, although some of them remained experimental only. The CLEAR evaluations used data from CHIL and AMI, as well as NIST and VACE (Chen *et al.*, 2005), some of it being shared with RT.

Beyond the established scientific events and scholarly journals which disseminate work on meeting analysis and access, the community has also created a new dedicated forum, the Machine Learning for Multimodal Interaction (MLMI) workshops, initiated more specifically by the AMI and IM2 consortia. Many of the research results gathered in this book were originally presented at MLMI

workshops.[2] Due to converging interests and complementarity, joint events between MLMI and the International Conference on Multimodal Interfaces (ICMI) were organized in 2009 and 2010. Following their success, the two series merged their advisory boards and decided to hold annual conferences under the name of International Conference on Multimodal Interaction.

1.4 Outline of the book

In order to design tools with the potential to unlock the business value contained in meetings, researchers in several related fields must collaborate. There are many places to find information about components like speech recognition that are the building blocks for the new technology. However, understanding the global picture requires a basic understanding of work from a wide range of disciplines, and help for developing that understanding is much harder to find. One particular challenge is in how to use what organizational and social psychologists know about human groups to determine user requirements and methods of testing technologies that users cannot really imagine yet. Another is in joining work on individual communication modalities like speech and gesture into a truly multimodal analysis of human interaction. While this book does not pretend to offer a fully integrated approach, the longevity of the collaborations between its authors has enabled many new connections and the feeling that it was possible to understand and achieve more by working together. One of the goals of this book is to pass on that understanding, making it easier for new researchers to move from their single disciplines into a rewarding and exciting area.

The book begins with something that underpins everything that follows: the data. Chapter 2 presents a hardware and software infrastructure for meeting data collection and annotation, initially designed for the comprehensive recording of four-person meetings held in instrumented meeting rooms. The rooms were used to record the AMI Meeting Corpus (Carletta, 2007), which consists of 100 hours of meeting recordings, along with manually produced transcriptions and other manual annotations that describe the behavior of meeting participants at a number of levels.

After Chapter 2, the book contains two unequal parts: Chapters 3–10 and Chapters 11–13. The first part explains the range of technological components that make up multimodal signal processing. Each chapter takes one kind of analysis that an application might need and describes what it does, how it works (and how well), and what the main issues are for using it. The advances in audio, visual, and multimodal signal processing are primarily concerned with the development of algorithms that can automatically answer, using the raw audio-video streams, questions such as the following ones: What has been said during the meeting? Who has spoken when? Who and where are the persons in the meeting? How do people behave in meetings? What is the essence of what has been said? In general, the order of the chapters reflects a progress towards

[2] The workshop proceedings were published as revised selected papers in Springer's Lecture Notes in Computer Science series, numbers 3361 (Martigny, 2004), 3869 (Edinburgh, 2005), 4299 (Bethesda, MD, 2006), 4892 (Brno, 2007) and 5237 (Utrecht, 2008).

more and more content abstraction, building up higher and higher levels of information from raw audio and video signals.

Chapters 3 to 5 build up towards an understanding of what was said in a meeting, primarily (but not entirely) based on audio signals, from microphone arrays (Chapter 3) to speaker diarization (determining who spoke when, Chapter 4) and automatic meeting transcription (Chapter 5). Chapters 6 and 7 move to focus more substantively on video processing as a source of information, again building upwards from the raw signals. Chapter 6 deals with tracking individual people, and especially their heads, as they move through a space. Chapter 7 then builds on this work to discuss methods for finding people and faces in recordings, recognizing faces, and interpreting head and hand gestures.

The remaining chapters in the first part of the book develop more of what a layperson would consider an understanding of a meeting. Chapter 8 describes analyses that begin to make sense of the words that were said, such as removing disfluencies, identifying questions, statements, and suggestions, or identifying subjective statements, such as positive opinions. Chapter 9 is more social in nature, and covers the analysis of conversational dynamics, in particular in terms of which speakers are being most dominant conversationally, and the different roles that they take in the meeting. Finally, Chapter 10 addresses a higher-level but very important task: that of creating useful summaries of meetings.

The second part of the book (Chapters 11–13) considers how to design, build, and test applications that use multimodal signal processing to analyze meetings. It takes the reader from the methods for identifying user needs for meeting support technology and their results (Chapter 11), through a range of meeting browsing applications that draw on underlying components from the first part (Chapter 12), to the methods for evaluating them (Chapter 13). The focus is particularly on meeting browsers, the most mature of the new technologies, which allow users to find information from past meetings, but the material also covers applications that support groups as they meet.

Finally, the conclusion (Chapter 14) abstracts from the lessons learned in analyzing meetings, and adopts a critical perspective to show what interesting and scientific challenges are still left ahead of us, and their potential impact in other application domains, such as social signal processing.

1.5 Summary and further reading

Multimodal signal processing has now had a decade of investment, including the promotion of shared tasks that allow the results from different techniques to be compared. It has benefited immensely from hardware advances that make synchronized recordings of audio and video signals relatively cheap to make and store. There are now many different automatic analyses available as components for systems that will do new, useful, and interesting things with these recordings. Although meeting support technology is only one of the many possibilities, the emergence of corporate meeting archives and the business value locked in them make it an obvious choice.

We conclude this introduction (and, indeed, every chapter of this book) with suggestions for further reading. These include mostly books at comparable levels of generality; more focused articles on specific topics are indicated in the respective chapters, while the names of relevant periodicals and conference series can simply be found by browsing the bibliography at the end of the book.

The books by Thiran *et al.* (2010) and Waibel and Stiefelhagen (2009) draw on some of the same core technologies as the present book, but cover certain additional aspects not dealt with here, such as human–computer interaction (HCI), speech synthesis, or multimodal fusion. The second book is a collection of papers summarizing achievements from the CHIL project, each of them with a close focus on specific research results. Books like those by Cassell *et al.* (2000) and by Stock and Zancanaro (2005) are in the same general area of multimodal interaction, but focus on presenting, not obtaining, information from multimodal data. An overview of machine learning algorithms for processing monomodal communication signals similar to those analyzed in this book is provided by Camastra and Vinciarelli (2008). There are many books about multimodal HCI, such as those by Wahlster (2006), from the SmartKom project, or by Grifoni (2009), which include spoken and multimodal dialogue interfaces and mobile devices. The proceedings of the MLMI conferences series of work mentioned in Section 1.3.4 represent additional collections of in-depth research articles (e.g., Popescu-Belis and Stiefelhagen, 2008).

1.6 Acknowledgments

Most of the contributors to this book, though not all, have been connected to some extent to the AMI Consortium. The editors and authors are grateful for the significant support of the European Union, through the Sixth Framework Programme for Research in its Information Society Technology (IST) thematic priority, as well as the support of the Swiss National Science Foundation through its NCCR division.

More specifically, the following grants have supported the research presented here, as well as the preparation of the book itself: the AMI EU integrated project (FP6, no. IST-2002-506811), the AMIDA integrated project of the EU (FP6, no. IST-033812), and the IM2 NCCR of the Swiss SNSF. Unless otherwise stated, the research work described in this book was funded by these sources. Additional funding sources are acknowledged at the end of each chapter.

The editors would like to thank the staff at Cambridge University Press, in particular Dr. Philip Meyler and Ms. Mia Balashova, their copy-editor Mr. Jon Billam, as well as Dr. Pierre Ferrez from Idiap, for their help with the production of this book.