View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Infoscience - École polytechnique fédérale de Lausanne

Chapter 1

Interactive multimodal information management: shaping the vision*

In the past twenty years, computers and networks have gained a prominent role in supporting human communications. This constitutes one of the most remarkable departures from their initial role as processors of large amounts of numeric data, in business or science, or as controllers of repetitive industrial operations. However, to offer truly innovative support to human communications, computers had to demonstrate that they could achieve more than what telephone calls or videoconferencing could do.

Thanks to research in the past decade, a convincing case has been made for the capability of information and communication technology to do justice to one of the richest aspects of human communication: its multimodal nature. Humans generate meaningful communicative actions with far more means than only pronouncing or writing words. Individuals words are put together in sentences and dialogues, accompanied by nuances of tone and pace, face expressions and gestures. Utterances from different speakers build together complex, multimodal interaction patterns which are some of the richest, yet natural social activities. This book is an attempt to answer the questions: what is required from computer hardware and software to support such activities? What are the capabilities of current technology, and what can be achieved using it?

This book takes a strong stance. We posit that research in interactive multimodal information management by computers makes quicker progress when it is driven by a clear application goal, which not only provides concrete

*This is a draft version of the following book chapter: Andrei Popescu-Belis and Hervé Bourlard, "Interactive multimodal Information Management: Shaping the Vision", in Bourlard H. and Popescu-Belis A., editors, *Interactive multimodal Information Management*, EPFL Press, Lausanne, 2013, pages 1-17.

use cases, and a sense of social and economic utility, but above all constitutes a controlled experimental framework which is essential to empirical science.

The framework put forward in this book is centered on the capture, automatic analysis, storage, and interactive retrieval of multimodal signals from human communication as it occurs in meetings. This framework has shaped the vision of the contributors to this book, and of many other researchers cited in it. In the past decade, this vision has opened an entirely new array of problems, by offering at the same time the much needed empirical data that is characteristic of recent research in information processing based on machine learning. Moreover, this framework has received significant long-term institutional support through an array of projects, reviewed below, and including the Swiss National Center of Competence in Research (NCCR) in Interactive Multimodal Information Management (IM2), to which all contributors of the book have been connected.

In this introduction, we sketch the overall concept of meeting capture, analysis and retrieval which forms the backbone of the book – although individual chapters emphasize the underlying research achievements rather than a particular system, and often go beyond the meeting framework. The hardware components and the data that are crucial to the proposed framework are presented, followed by a brief historical and organizational tour of the IM2 NCCR and a review of related projects.

1.1 Meeting capture, analysis and access

Managing multimodal information to support human communication can take many forms, but progress is best achieved when this highly multifaceted effort can be coordinated under the scope of a common vision and application. One of the most fruitful approaches of the past decade has been centered around the concept of meeting support technology, which underlies most of the studies presented in this book. This vision answers the challenge of finding a concrete setting to drive research by making available large amounts of benchmark data and ensuring that multimodal analysis and delivery processes are piped together in a coherent fashion. In this section, we outline the main building blocks of meeting support technology as they appear in this book.

1.1.1 Development of meeting support technology

Among various ways of enhancing human meetings through information technology, we focus in this book on offline support under the form of meeting capture, analysis, and retrieval systems. These can be seen as intelligent archival and access systems for multimodal data from meetings, intended either for people who attended a given meeting and would like to review it or check specific information at later moments, or for people who missed a

meeting and would like to obtain a digest of what happened in it. Moreover, support for accessing series of meetings is part of the overall scenario as well.

Considering such an application inevitably raises the question of the respective importance that should be given to users' needs versus researchers' interests in shaping the intended application scenario. One point of view is that research and development should start only when users' needs have been properly assessed, in other words, only when developers have fully understood how the technology would be used to enhance people's efficiency at preparing, attending, and reviewing meetings. Another point of view argues that users might not be aware of actual or future capabilities offered by meeting support technology, and that some of their desiderata might even be biased by misconceptions about technology. According to this view, researchers and technology providers should have control over the most promising R&D directions to explore, with the risk of low market uptake or irrelevance to users. In this book, rather than adopting one of these points of view, most of the chapters take an intermediate position, which incorporates indicators of potential relevance to users in the decisions made to address specific research problems. Research-driven chapters are thus central, but most of them include an application-oriented component. Therefore, the development process often loops several times through the specification-implementation-evaluation cycle.

1.1.2 Scenario and context

Several classifications can be applied to meeting support technology. For instance, it is possible to distinguish between *online* support *during* meetings, e.g. to improve the discussion flow or help with document retrieval, and *offline* support *between* meetings, e.g. to help writing minutes or accessing past content. Another possible distinction concerns the type of group interactions (Bales, 1950, McGrath, 1984), which can be categorized in terms of the setting (business or private), the number of participants (two people, small group, large group), the form of interaction (discussion or presentation with questions), and the purpose (brainstorming, decision making, or problem solving).

The techniques put forward in this book, for unimodal and multimodal processing, multimedia retrieval, and human-computer interfaces are applicable to a large range of settings. However, the book focuses on the federating scenario, supported by shared data and common targets in the IM2 NCCR, of meeting capture, analysis, storage and retrieval. More specifically, the unifying vision of IM2 is centered around small-group professional meetings, with discussions but also presentations, which are held in smart meeting rooms equipped with sensors for several human communication modalities: audio, video, handwriting, and slide projection. The global target of interactive multimodal information management, in this context,

is the analysis of human communicative signals in order to extract meaningful features that are used for the indexing and retrieval, through dedicated interfaces, of the information content of a human meeting. This target has been shared by the IM2 NCCR with several other large initiatives (see Section 1.3) and is a prominent instance of the *interaction capture and retrieval* scenario described by Whittaker et al. (2008). The chapters gathered in this book can be seen as a coherent picture of what technology can offer today for reaching such an ambitious goal.

The main functionalities of a meeting capture, archival and retrieval system are the following:

1. Capture of human communicative signals in several modalities, thanks to dedicated hardware situated in instrumented spaces such as the smart meeting rooms presented in Section 1.1.3 below.
2. Analysis of human communicative signals to automatically extract meaningful characteristics such as words (from the audio signals) or face expressions (from the video signals). These analyses can be first performed for each communication modality in part, through speech and language processing (see Part III, Chapters 15 through 18), image and video processing (see Part II, Chapters 8 through 11) and document processing (Chapters 19, 20 and 21). Additional valuable information can be extracted from multimodal processing, which is crucial for understanding human interactions in meetings (Chapters 12 and 13 present two prominent examples). While these analyses are mainly aimed at extracting features for retrieval, they can also be used for event detection, meeting summarization, or person recognition (see Chapters 8 and 13).
3. Storage and access, which are briefly discussed in Section 1.1.4.
4. Interactive meeting search or meeting browsing (see Part I, Chapters 4, 5, and 6), including multimedia information retrieval (Chapter 7) and possibly meeting assistants (Chapters 2 and 6).

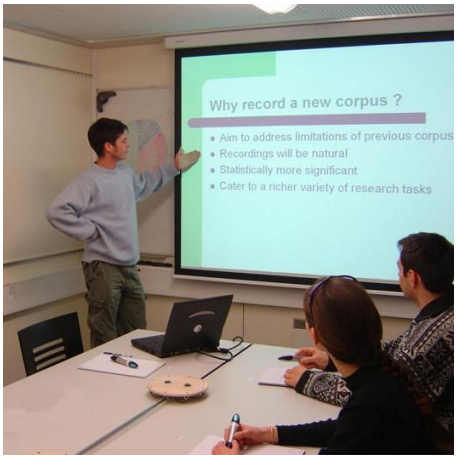
In addition, several transversal elements are equally important to the meeting capture and retrieval scenario, and are represented in some of the chapters of this book:

1. User studies and human factors, mainly discussed in Part I.
2. Evaluation protocols, either for the entire meeting capture and retrieval application (Chapter 6) or as Wizard-of-Oz evaluations of interfaces (Chapter 4).
3. Data collection and the resulting resources, presented in the next section (1.1.4).

1.1.3 Smart meeting rooms

The main vision adopted in this book is that research on interactive multimodal information management requires in the first place a range of devices to capture multimodal signals from human communications, such as speech, gestures or documents. To ensure fair comparisons between studies and to facilitate the integration across modalities, the nature and placement of the capture devices in the physical space must be precisely defined and kept constant across studies. In the context of this book, the instrumented spaces in which meeting data capture took place are known as *smart meeting rooms (SMRs)*.

One of the first fully specified SMRs, shown in Figure 1.1, was built in 2001–2002 at the Idiap Research Institute (Moore, 2002), at the same time as several similar attempts at other institutions (e.g. Chiu et al., 2001, Cutler et al., 2002, Lee et al., 2002, Stanford et al., 2003). The Idiap SMR, described below, was used to record a large amount of data. Moreover, it was reproduced with nearly identical settings at two other institutions, the University of Edinburgh and TNO in the Netherlands. These three SMRs were all used to record the AMI Meeting Corpus presented below (Carletta et al., 2005a, Carletta, 2007). Another SMR was setup at the University of Fribourg (Lalanne et al., 2003), with different specifications, accommodating a larger number of meeting participants than the Idiap SMR (see Figure 1.2).



(a) Prototype of the Idiap SMR. View towards the projection screen, with a presenter and two participants.

(b) Permanent state of the Idiap SMR. Note the microphone array at the center of the table, with face-capture cameras below it, and a wide-field camera above the books.

Figure 1.1: Smart meeting room (SMR) at the Idiap Research Institute, for up to four participants in the displayed configuration (Moore, 2002).

The precise description and placement of capture devices in a SMR ensures that geometric and electronic specifications are fully known when performing unimodal and multimodal signal processing, and that methods can be compared in exactly the same conditions. Detailed descriptions of the Idiap and University of Edinburgh SMRs can be found elsewhere (see Moore, 2002, Renals et al., 2012, Chapter 2). For the general understanding of the research presented in this book we provide here a brief outline.

Figure 1.2: Smart meeting room at the University of Fribourg, for up to eight participants (Lalanne et al., 2003). Note the hemispheric casings of the individual cameras and the table-top individual microphones.

The Idiap SMR has in the first place the functional characteristics of a standard meeting room. The large table can seat up to 12 people, but in most of the configurations there were only four participants: two persons sitting on each of the opposite sides of a rectangular table, as shown in Figure 1.1. The equipment includes a video projector (beamer) with a dedicated projection screen, a white board, as well as the possibility to use individual laptops (with Internet and individual beamer connections) and to make notes on paper. The room is isolated from external noise and has fluorescent lighting rather than windows, to ensure uniform, low-noise audio-visual recording conditions.

The capture devices were designed to be as non-intrusive as possible, to preserve the naturalness of the interaction. They included three types of microphones: head-mounted and lapel individual microphones, along with a central, circular, 8-channel microphone array (visible in Figure 1.1b). These three sorts of microphones range from the most to the least intrusive ones, but also, conversely, from the most to the least accurate ones. Capturing

audio on several channels raises the possibility to reconstruct as closely as possible the audio signals emitted by participants (speech and noises). For instance, the output of a microphone array can be compared to the reference recordings from each of the head-mounted microphones.[1] The Idiap SMR thus supports capture and recording of up to 24 audio channels – two per participant and one or two 8-channel arrays – which are digitized and streamed directly to a computer hard disk, using three 8-channel PreSonus Digimax pre-amplifiers / digitizers and a PC audio interface.

The Idiap SMR also records up to eight video streams: (a) three wide-angle cameras offering a view of each of the two sides of the table and the front of the room; (b) four individual cameras positioned under the microphone array; and (c) the RGB signal from the video projector, capturing exactly what participants see on the projection screen, generally images of slides. The capacity to capture the screen of each participant's laptop was added at a later stage. Initially, the video signals from the wide-angle cameras were recorded using MiniDV technology, and later using multiple-channel Firewire video acquisition cards.

To enable the study of modality fusion, the audio and video systems are synchronized using a master sync signal, each channel being accurately time-stamped. All cameras are frame-locked using a master black burst synchronization signal from a Horita BSG-50 device. A time code that is also synchronized with the master one is generated using a MOTU MIDI Timepiece AV timing control module and added to the audio and video recordings. The 48 kHz clock used for audio digitization is also derived from the master sync signal. Only the handwriting information, capture using Anoto pen technology, was not synchronized at the same level of precision, as it relied only on the pens' internal timing information.

1.1.4 Data: multimodal signals and their annotations

The smart meeting room infrastructure allows the capture of several modalities that support human communication: videos of the room or videos focused on faces, audio signals from human voices, but also drawings on a white board, notes taken on paper, and documents that are projected. Additional material related to a meeting, not necessarily presented during the meeting, can be added to these recordings if it is available in digital form. The recording and storage of all these signals provides the raw material upon which most of the research presented in this book is built.

However, recordings of raw signals are of little utility without accompa-

[1] Research on microphone arrays has played a central role in audio and multimodal processing for meetings, as described for instance by McCowan (2012). This research started from initial experiments with various numbers and configurations of microphones, and reached the stage of commercial products: the Microcone, now available at Apple Stores, is an example of successful technology transfer (along with others listed in Chapter 23).

nying *annotations*, that is, additional indications of meaningful interpretation units over the signals. The paradigm on which most multimodal processing research is built is that the output of the processing modules can be represented as *automatic annotations* of the signals. For instance, the exact words uttered by a meeting participant, together with the exact time when they were pronounced, constitute an annotation (commonly called 'speech transcript'). Automatic speech recognition systems (see Chapter 15) aim at finding this information automatically, from a more or less noisy audio signal, and are evaluated in terms of accuracy with respect to a true transcript produced by a human, called 'ground truth' or 'gold standard'. Annotations of time-dependent signals can take several forms, abstractly represented as segmentation, labeling of segments, relating the segments, labeling the relations, and so on.

Metadata and *annotations* are related notions, as they both refer to additional information about meaningful items in a raw signal. Throughout this book, 'annotations' will refer to time-dependent information, while 'metadata' will characterize a recording in its entirety (though this distinction is not universal). For instance, segmenting audio into speaker turns is an annotation (see Chapters 16 and 17), but listing the participants to a meeting is part of the metadata (see e.g. Chapters 8 and 13).

To estimate the accuracy of the output produced by a unimodal or multimodal processor, it must be compared to a *reference annotation* of the same signal, generally produced by human annotators (see e.g. Pustejovsky and Stubbs, 2012, for language annotations). The availability of ground truth annotations is essential for two purposes: (1) to evaluate processing software by comparing their results to the desired ones; (2) to train software using machine learning methods, i.e. by learning the correspondences between features of an input signal and the desired annotation.

The Idiap SMR, along with its two clones at the University of Edinburgh and TNO, has served to record the AMI Meeting Corpus, subsidized by several projects including the IM2 NCCR. This freely available corpus comprises more than 100 hours of scenario-based and free-form meetings, with recordings of all the modalities listed in the above section (Carletta et al., 2005a, Carletta, 2007, Renals et al., 2012, Chapter 2). A scenario based on series of four meetings that were aimed at the design of a remote control for TV sets was defined to allow studies of group interaction in controlled conditions. The most remarkable feature of the AMI corpus is the extent of the manual annotations that were made, often covering the entire data.

The following modalities and dimensions have been annotated: speech segmentation, word-level transcription (with forced time alignment between correct words and audio signals), named entities, dialogue acts, topics, summaries, head and hand gestures, gaze direction, and movement around the room. Most of the annotations were done with the NITE XML Toolkit

(NXT) (Carletta et al., 2003, 2005b), and all of them are now distributed in this format. The format can easily be converted to the input/output formats of other processing tools, and is also made available in a database structure (Popescu-Belis and Estrella, 2007).[2]

A storage and distribution infrastructure is needed to support the dissemination of multimodal data and annotations. The MMM server (see `www.idiap.ch/mmm/`) developed at the Idiap Research Institute is such a platform, now complemented by the Idiap Dataset Distribution Portal. Two front-ends have been created for the AMI corpus: one for the data described above (`http://corpus.amiproject.org`), and another one for the AMIDA data which includes a distant participant to the meetings (`http://corpus.amidaproject.org`). Descriptive metadata was created for the AMI corpus in OLAC format (Open Language Archives Community, derived from the Open Archives Initiative), which integrates the data into a large catalog of language and multimodal resources accessible via the OLAC metadata harvesting engine (`www.language-archives.org`). Moreover, when it comes to producing and accessing annotations in real time, using multimodal processing software, a client/server architecture based on generic annotation triples was designed (the Hub, `www.idiap.ch/mmm/tools/hub/`).

1.2 The IM2 Swiss National Center of Competence in Research

We have shown above how the general topic of interactive multimodal information management (in short, IM2) could be translated into a concrete research scenario, namely meeting capture, analysis and retrieval. This vision was shaped within the IM2 National Center of Competence in Research (NCCR), which has gathered a large number of research institutions active in the above-mentioned research fields, over a period of twelve years. In this section, we will briefly review the history and structure of the IM2 NCCR, to show how the initial vision was put into practice through the management of science.

1.2.1 History of IM2

The NCCR concept as a long-term research funding instrument was presented by the Swiss National Science Foundation (SNSF) in August 1998. It was approved by the political authority, the Swiss Federal Council, in November of the same year. A first call for declarations of intent from

[2]Other multimodal corpora for the study of interactive multimodal information management exist (Kipp et al., 2009), but none is as extensive as the AMI corpus. The CHIL corpus contains multimodal recordings of lectures with several annotations (Mostefa et al., 2007), especially non-verbal ones, while many other conversational corpora, including the ICSI Meeting Recorder corpus (Janin et al., 2003) are limited to the audio modality.

all Swiss research institutions was issued in January 1999, and the outline of IM2 was proposed among them. The idea of conducting coordinated research in multimodal information processing and interactive access was thus conceived in early 1999. The proposed leading house was the Idiap Research Institute in Martigny, a young and independent organization, in contrast to most of the other ideas coming from established institutions of higher education in Switzerland.

No fewer than 230 declarations of intent were received by the SNSF in March 1999. However, turning them into pre-proposals was quite a challenge, as only 82 pre-proposals were submitted in July 1999 for scientific evaluation by an international panel. As has since become the norm, the selection process rated the pre-proposals into three categories, based on the expected chances of success: IM2 was included, along with 27 others, in the first one. Although all consortia were allowed to submit a full proposal, following pre-screening only 33 full proposals were submitted in March 2000. An extensive scientific and political evaluation took place, including an oral presentation for 18 selected proposals.

The decisions were announced in December 2000 by the Federal Council: ten NCCRs were to receive immediate funding, while four more were to be supported on condition that the Swiss Parliament approved additional funds for the program, which was done in June 2001. Therefore, after more than two years of gestation, the IM2 NCCR was born for good[3]. Following an initial planning and recruitment period, the official starting date of IM2 was January 1st, 2002. As a leading house, this initiated a period of significant growth for Idiap, under the direction of the first editor of this book, also the head of IM2. Moreover, the relation of Idiap to EPFL was strengthened by the nomination of an IM2 deputy director from EPFL.

1.2.2 Size and management of IM2

As defined by the SNSF in the general regulations of NCCRs, funding was allocated for at most three periods of four years, called phases, conditioned on satisfactory annual scientific reviews and on the approval of full scientific proposals for renewal between phases (in 2005 and 2009). The transitions between phases have been accompanied by substantial evolutions in the IM2 structure, to match more effectively the effort intended to achieve its vision. With a total subsidy from the SNSF of about 32 million Swiss francs (CHF) for twelve years, the IM2 members leveraged additional, non-SNSF funding from European projects (such as AMI or AMIDA mentioned below) or from industrial collaborations, as well as their own institutional funding, to reach a total budget of nearly 85 million CHF over twelve years. Due to the different strategic profiles of the three phases, the yearly IM2 budget was

[3]Since then, calls for NCCRs have been issued every four years, leading to five new NCCRs in 2005, eight in 2010, and about five expected for 2014.

about 7.5 million CHF in phases I and II (2002–2009), but only about half of this amount in phase III (2010–2013).

Six major institutions have taken part in the IM2 NCCR. Along with the leading house, Idiap, these were the Universities of Geneva, Fribourg, and Bern (until 2010), and the two Federal Institutes of Technology in Lausanne and Zurich (EPFL and ETHZ). Several other institutions have been involved for variable periods of time in IM2: the HES in Sion and Fribourg, the CSEM in Neuchâtel, and the International Computer Science Institute (ICSI, Berkeley) for the first two phases, including a successful student exchange program. From each institution, several teams or labs have participated, with an annual average of about 20 labs in the first two phases, and 10 in the third one (not counting Idiap's research groups individually).

A large number of researchers have been involved in IM2, i.e. they were at least partially subsidized by the SNSF IM2 grant or by matching funds. In the first two phases, 150–200 people have contributed yearly to IM2, going below 100 only at the beginning of the third phase. IM2 supported mainly doctoral students, besides postdocs, researchers and professors: there were about 50 to 70 PhD students in any given year, more than half of them from Idiap, gradually decreasing towards the end as theses were defended and fewer students were hired.

The steering and coordination of the IM2 NCCR was ensured by a strong organization. The director and deputy director worked in close connection with a Technical Committee (TC) comprising the heads of all individual projects. In the second phase, the TC was renewed with more junior members, as an opportunity for them to increase their decisional abilities, and to make the TC more closely related to day-to-day research. In parallel, a Steering Committee was created to include one senior member from each participating institution. In the third phase, a General Assembly involving all IM2 group leaders replaced the two committees.

Invaluable feedback was obtained from the SNSF-appointed Review Panel, which met every year with IM2 representatives to evaluate their progress, to make recommendations for future work, and to decide the continuation towards the second and then the third phase. The Scientific Advisory Board, appointed by the IM2 management, has issued advice at yearly meetings with IM2. Several members of these boards have kindly accepted to be interviewed to provide brief assessments of IM2, gathered in Chapter 22.

To ensure communication and coordination among all its members, IM2 has organized an annual series of summer institutes, featuring talks from IM2 members (with emphasis on PhD work), from invited speakers, as well as a variety of panel discussions, training sessions, and social activities – for up to 100 participants every year. The following events have been organized:

1. IM2 Summer Institute, Martigny, October 3–4, 2002.
2. IM2 Summer Institute, Crans-Montana, October 6–8, 2003.

3. Joint event with IM2 sessions at the first MLMI workshop (see Section 1.3), Martigny, June 21–23, 2004.
4. IM2 PhD Integration Week, Moudon, August 16–18, 2004.
5. IM2 Summer Institute, Lausanne, November 14-17, 2005.
6. IM2 Vision Day, Geneva, September 3-4, 2006.
7. IM2 Winter Institute, Löwenberg Center, Murten/Morat, February 19–22, 2007.
8. Joint IM2 and Affective Sciences NCCR Summer Institute, Riederalp, September 1–3, 2008. The two NCCRs have collaborated, since then, on issues related to non verbal communication and social signals.
9. IM2 Summer Institute, Chavannes-de-Bogis, August 31–September 2, 2009.
10. IM2 Summer Institute, Saanenmöser, September 13–15, 2010.
11. Joint event with IM2 sessions at Idiap's 20th anniversary celebration, Martigny, September 1–2, 2011.
12. IM2 Summer Institute within the International Create Challenge (see Chapter 23), Martigny, September 3–4, 2012.
13. IM2 Final Event, Lausanne, October 17-18, 2013, at which this book will be launched.

1.2.3 Structure of IM2

The structure of the NCCR was based on individual projects or IPs. But 'individual' did not mean that they concerned individual persons, or that they made progress separately from each other. In fact, each IM2 IP grouped several partners working closely together on the same problem, with many connections being made across IPs as well. The structure has evolved from one phase to another, reflecting variations in focus, though always globally paralleling the tasks necessary to achieve the IM2 goals. Moreover, at various stages of the NCCR, internal calls for 'white papers' or 'mini-projects' have ensured that the most urgent tasks received additional support when needed.

The *phase I structure (2002–2005)* included the following IPs, presented here in the order that matches most closely the IM2 vision:

1. **SA:** Scene Analysis, computer vision research on image segmentation, face analysis, and handwriting recognition.
2. **SP:** Speech Processing, on speech segmentation, recognition, and synthesis.
3. **ACP:** Multimedia Information Access and Content Protection, on biometric features and person identification.

4. **DI:** Document Integration, bridging the gap between non-temporal documents and other temporal media.
5. **MI:** Multimodal Input and Modality Integration, on the fusion and decoding of several input modalities.
6. **MDM:** Multimodal Dialogue Management, on dialogue modeling for human-human and human-computer dialogue.
7. **DS:** Deployment, Storage and Access to Multimodal Information, on multimedia databases, later merged with the Integration Project.
8. **IIR:** Information Indexing and Retrieval, with emphasis on multimedia search.
9. **IP:** Integration Project, with support for system integration across IPs.

The *phase II structure (2006–2009)* included the following IPs, some of which carried over quite clearly certain research fields from phase I, while others were new:

1. **DMA:** Database Management and Meeting Analysis.
2. **AP:** Audio Processing.
3. **VP:** Visual/Video Processing.
4. **MPR:** Multimodal Processing and Recognition.
5. **MCA:** Multimodal Content Abstraction.
6. **HMI:** Human-Machine Interaction.
7. **ISD:** Integration Software and Demonstrators, integrated into DMA at the middle of phase II.
8. **BMI:** Brain Machine Interfaces, as the EEG modality appeared of particular interest at the end of phase I.

The *phase III structure (2010–2013)* was intended to accommodate the phasing-out of the SNSF financial support and consolidate the IM2 achievements, in preparation for the post-IM2 period, while at the same time applying the theoretical and practical results of the first two phases in a new environment (interactions in educational settings) and focusing on the particularly promising field of social signal processing (Gatica-Perez, 2009, Vinciarelli et al., 2009). The phasing-out structure consisted of only three highly-integrated individual projects.

1. **IP1:** Integrated Multimodal Processing, pursuing the most important research directions from the first two phases, with an integration and evaluation component.

2. **IP2:** Human Centered Design and Evaluation, aiming to generalize the IM2 technologies by applying them to new environments, other than smart meeting rooms, in combination with third-party technology, and testing their acceptance by various groups of users, particularly in educational settings.
3. **IP3:** Social Signal Processing, understanding of social signals through automatic analysis of nonverbal communication, and applying these approaches to meeting analysis.

The AIM2 association, created in 2012, will continue the activities of the IM2 Consortium after 2013 under a different operating and funding structure. More details about it are given in Chapter 23 of this book, dedicated to technology transfer.

1.3 Related international projects and consortia

Several large initiatives focusing on multimodal signal processing and its application to meeting support technology have been contemporary to IM2. These international projects have been related to IM2 due to their similarity of topic, but also more concretely through the participation of one or more IM2 members into these consortia.

During the 1990s, advances in the audio and video analysis of recordings have lead to the first implemented systems for interaction capture, analysis and retrieval (e.g. Whittaker et al., 1994, Kubala et al., 1999). But the first project to apply multi-channel audio recording and processing to business meetings was the Meeting Recorder project at ICSI, Berkeley (Morgan et al., 2001, 2003), which produced a landmark corpus that was reused in IM2. Around the year 2000, it became apparent that technologies for meeting support needed to address a significant subset of the modalities used for human communication, using appropriate capture devices in instrumented meeting rooms (Chiu et al., 2001, Cutler et al., 2002, Lee et al., 2002, Stanford et al., 2003).

The need for advanced multimodal signal processing applied for meetings and lectures was addressed in the past decade by several consortia doing mainly fundamental research, briefly presented below[4]. Three main groups of consortia can be identified, considered in a large and non-disjoint sense, because many collaborations across projects, e.g. for data sharing, have taken place. The first group gathered around CMU/ISL and the University of Aachen, with the FAME and CHIL projects, emphasizing lectures, video processing and event detection. In the US, a second group evolved from the ICSI MR to the CALO project, with emphasis on language and semantic

[4]See also the books by Waibel and Stiefelhagen (2009), Thiran et al. (2010), and Renals et al. (2012).

analysis. The third one, around the Idiap Research Institute and the University of Edinburgh, evolved from the M4 to the AMI/AMIDA projects, related to IM2, with a wider approach including unimodal, multimodal, and semantic analyses.

CMU's Interactive Systems Laboratory initiated a project on meeting record creation and access (Waibel et al., 2001), while the FAME European project (Facilitating Agent for Multicultural Exchange, 2002–2005) developed a prototype for using multimodal information streams in an instrumented room (Rogina and Schaaf, 2002, Metze and al., 2006). 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 (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.

The M4 European project (MultiModal Meeting Manager, 2002–2005) achieved a complete system for multimodal recording, structuring, browsing and querying meetings (McCowan et al., 2003, 2005). Then, the AMI Consortium (EU project on Augmented Multiparty Interaction (AMI), 2003–2006, later with distance access (AMIDA), 2006–2009) carried out research in meeting analysis and technology on a large scale, with a wide focus on multimodal signal processing, meeting summarization and browsing, and human factors and evaluation (Renals et al., 2012). IM2 was strongly related to the AMI consortium, as Idiap served as a leading house for both projects, in collaboration with the University of Edinburgh for AMI.

Beyond the established scientific events and scholarly journals which disseminate work on meeting analysis and access, these communities have created a new dedicated forum, the Machine Learning for Multimodal Interaction (MLMI) series of workshops, started in 2004. 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 new name of International Conference on Multimodal Interaction.

Acknowledgments

The contributors to this book are members of the IM2 NCCR. Unless otherwise stated, the research work described in this book was funded by the

IM2 NCCR. The editors and authors are thus very grateful for the significant long-term support of the Swiss National Science Foundation through its NCCR Division. The two editors would also like to thank the staff at EPFL Press for their kind assistance during the publication process.

Bibliography

Bales, R. F. (1950). *Interaction process analysis: A method for the study of small groups*. Addison-Wesley, Reading, MA, USA.

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

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2005a). The AMI Meeting Corpus: A pre-announcement. In *Proceedings of the 2nd International Workshop on Machine Learning for Multimodal Interaction (MLMI 2005)*, pages 28–39, Edinburgh, UK.

Carletta, J., Evert, S., Heid, U., Kilgour, J., and Chen, Y. (2005b). The NITE XML Toolkit: Data model and query language. *Language Resources and Evaluation*, 39(4):313–334.

Carletta, J., Evert, S., Heid, U., Kilgour, J., Robertson, J., and Voormann, H. (2003). The NITE XML Toolkit: Flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers*, 35(3):353–363.

Chiu, P., Boreczky, J., Girgensohn, A., and Kimber, D. (2001). LiteMinutes: an internet-based system for multimedia meeting minutes. In *Proceedings of the 10th international conference on World Wide Web (WWW 2001)*, pages 140–149, Hong Kong, CN.

Cutler, R., Rui, Y., Gupta, A., Cadiz, J. J., Tashev, I., He, L., Colburn, A., Zhang, Z., Liu, Z., and Silverberg, S. (2002). Distributed Meetings: A meeting capture and broadcasting system. In *Proceedings of the 10th ACM International Conference on Multimedia (ACM Multimedia 2002)*, pages 503–512, Juan-les-Pins, FR.

Gatica-Perez, D. (2009). Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing, Special Issue on Human Naturalistic Behavior*, 27(12):1775–1787.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). The ICSI Meeting Corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, pages 364–367, Hong Kong, CN.

Kipp, M., Martin, J.-C., Paggio, P., and Heylen, D. (2009). *Multimodal corpora: from models of natural interaction to systems and applications*, volume 5509 of *LNCS*. Springer-Verlag, Berlin/Heidelberg.

Kubala, F., Colbath, S., Liu, D., and Makhoul, J. (1999). Rough'n'Ready: a meeting recorder and browser. *ACM Computing Surveys*, 31(2es):7.

Lalanne, D., Sire, S., Ingold, R., Behera, A., Mekhaldi, D., and Rotz, D. (2003). A research agenda for assessing the utility of document annotations in multimedia databases of meeting recordings. In *Proceedings of 3rd International Workshop on Multimedia Data and Document Engineering*, Berlin, DE.

Lee, D., Erol, B., Graham, J., Hull, J. J., and N., M. (2002). Portable meeting recorder. In *Proceedings of the 10th ACM International Conference on Multimedia (ACM Multimedia 2002)*, pages 493–502, Juan-les-Pins, FR.

McCowan, I. (2012). Microphone arrays and beamforming. In Renals, S., Bourlard, H., Carletta, J., and Popescu-Belis, A., editors, *Multimodal Signal Processing: Human Interactions in Meetings*, pages 28–39. Cambridge University Press, Cambridge, UK.

McCowan, I., Bengio, S., Gatica-Perez, D., Lathoud, G., Monay, F., Moore, D., Wellner, P., and Bourlard, H. (2003). Modeling human interactions in meetings. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, pages 748–751, Hong-Kong, CN.

McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., and Zhang, D. (2005). Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):305–317.

McGrath, J. E. (1984). *Groups: Interaction and Performance*. Prentice-Hall, Englewood Cliffs, NJ, USA.

Metze, F. and al. (2006). The 'Fame' interactive space. In *Proceedings of Machine Learning for Multimodal Interaction (MLMI 2005)*, pages 126–137, Edinburgh, UK.

Moore, D. C. (2002). The Idiap Smart Meeting Room. Idiap Com 02-07, Idiap Research Institute.

Morgan, N., Baron, D., Bhagat, S., Carvey, H., Dhillon, R., Edwards, J., Gelbart, D., Janin, A., Krupski, A., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). Meetings about meetings: research at ICSI on speech in multiparty conversations. In *Proceedings of the*

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), pages 740–743, Hong Kong, CN.

Morgan, N., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Janin, A., Pfau, T., Shriberg, E., and Stolcke, A. (2001). The Meeting Project at ICSI. In *Proceedings of the 1st International Conference on Human Language Technology Research (HLT 2001)*, pages 1–7, San Diego, CA, USA.

Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S. M., Tyagi, A., Casas, J. R., Turmo, J., Cristoforetti, L., Tobia, F., Pnevmatikakis, A., Mylonakis, V., Talantzis, F., Burger, S., Stiefelhagen, R., Bernardin, K., and Rochet, C. (2007). The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language Resources and Evaluation*, 41(3-4):389–407.

Popescu-Belis, A. and Estrella, P. (2007). Generating usable formats for metadata and annotations in a large meeting corpus. In *Proceedings of the 45th Int. Conf. of the Association for Computational Linguistics (ACL 2007), Poster Sessions*, pages 93–96, Prague, Czech Republic.

Pustejovsky, J. and Stubbs, A. (2012). *Natural Language Annotation and Machine Learning*. O'Reilly Publishers, Sebastopol, CA, USA.

Renals, S., Bourlard, H., Carletta, J., and Popescu-Belis, A. (2012). *Multimodal Signal Processing: Human Interactions in Meetings*. Cambridge University Press, Cambridge, UK.

Rogina, I. and Schaaf, T. (2002). Lecture and presentation tracking in an intelligent room. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI 2002)*, pages 47–52, Pittsburgh, PA, USA.

Stanford, V., Garofolo, J., Galibert, O., Michel, M., and Laprun, C. (2003). The NIST Smart Space and Meeting Room projects: signals, acquisition annotation, and metrics. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, pages 736–739, Hong-Kong, CN.

Thiran, J.-P., Marqués, F., and Bourlard, H. (2010). *Multimodal Signal Processing: Theory and Applications for Human-Computer Interaction*. Academic Press, San Diego, CA, USA.

Tür, G., Stolcke, A., Voss, L., Peters, S., Hakkani-Tür, D., Dowding, J., Favre, B., Fernández, R., Frampton, M., Frandsen, M., Frederickson, C., Graciarena, M., Kintzing, D., Leveque, K., Mason, S., Niekrasz, J., Purver, M., Riedhammer, K., Shriberg, E., Tien, J., Vergyri, D., and Yang, F. (2010). The CALO Meeting Assistant system. *IEEE Transactions on Audio, Speech and Language Processing*, 18(6):1601–1611.

Vinciarelli, A., Pantic, M., and Bourland, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759.

Waibel, A., Bett, M., Metze, F., Ries, K., Schaaf, T., Schultz, T., Soltau, H., Yu, H., and Zechner, K. (2001). Advances in automatic meeting record creation and access. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001)*, pages 597–600, Salt Lake City, UT, USA.

Waibel, A. and Stiefelhagen, R. (2009). *Computers in the Human Interaction Loop*. Springer-Verlag, Berlin, DE.

Whittaker, S., Hyland, P., and Wiley, M. (1994). Filochat: Handwritten notes provide access to recorded conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence (CHI 1994)*, pages 271–277, Boston, MA, USA.

Whittaker, S., Tucker, S., Swampillai, K., and Laban, R. (2008). Design and evaluation of systems to support interaction capture and retrieval. *Personal and Ubiquitous Computing*, 12(3):197–221.