Extracting Information from Multimedia Meeting Collections

Daniel Gatica-Perez [1] Dong Zhang [1]
Samy Bengio [1]

IDIAP–RR 05-50

September. 2005

to appear in
7th ACM SIGMM International Workshop on Multimedia Information Retrieval

[1] IDIAP Research Institute, CP 592, 1920 Martigny, Switzerland, {gatica, zhang, bengio}@idiap.ch

IDIAP Research Report 05-50

Extracting Information from Multimedia Meeting Collections

Daniel Gatica-Perez Dong Zhang Samy Bengio

September. 2005

to appear in
7th ACM SIGMM International Workshop on Multimedia Information Retrieval

Abstract

Multimedia meeting collections, composed of unedited audio and video streams, handwritten notes, slides, and electronic documents that jointly constitute a raw record of complex human interaction processes in the workplace, have attracted interest due to the increasing feasibility of recording them in large quantities, by the opportunities for information access and retrieval applications derived from the automatic extraction of relevant meeting information, and by the challenges that the extraction of semantic information from real human activities entails. In this paper, we present a succint overview of recent approaches in this field, largely influenced by our own experiences. We first review some of the existing and potential needs for users of multimedia meeting information systems. We then summarize recent work on various research areas addressing some of these requirements. In more detail, we describe our work on automatic analysis of human interaction patterns from audio-visual sensors, discussing open issues in this domain.

1 Introduction

The value of recording, analyzing, accessing, and delivering multimedia meetings is manifold, as the number of existing research projects and commercial attempts seem to testify. Very broadly speaking, one can identify three different -fortunately not exclusive- positions regarding what is relevant about meeting collections: (1) what they *are*; (2) what *can* be done with them; and (3) what *needs* to be done with them. The first view acknowledges that meetings by themselves are relevant insofar as they constitute an expression of human interaction, the field of study of more than one branch of science [6, 25]. The second view regards meetings as an application domain where a diverse range of existing media technologies, including audio, speech, language, vision, information retrieval (IR), and human-computer interaction, can be tested and advanced [40, 28]. The third view considers meetings as a rich source of information with specific users and real needs to be satisfied [23, 18, 42], and where technology for meeting analysis is relevant as long as it addresses and contributes to satisfy user needs.

This paper starts from the last view, and aims at providing the unfamiliar reader with a concise introduction to this rapidly growing domain. We use the term "extracting information" in a sense than differs from the traditional text IR definition. Rather than a comprehensive review of existing work, we opt for a rapid description of what we believe to be research areas directed towards satisfying user needs, with up-to-date pointers to the literature. Our views are clearly influenced by our own work, which we review in more detail in later sections, in the context of recent and current research projects on the subject [51, 50, 49]. A recent overview of meeting technologies, with different emphasis to the one here, appears in [13].

The paper is organized as follows. Section 2 reviews existing work on user requirements. Section 3 summarizes relevant tasks and work in the various directions of the field. Section 4 presents various aspects of our recent work, discussing a number of open issues. Section 5 provides some concluding remarks.

2 What to extract from meetings: user requirements

Various studies have been conducted to distinguish potential users of multimedia meeting information systems, and to identify use cases and user requirements based on their information needs. Previous studies of user needs with a multimedia recording system in mind (any combination of audio, video, handwritten notes, electronic documents, etc. in a single system) can be traced back to [41, 27], and more recently to [23, 18, 42, 11]. In this section, we discuss user requirements for meetings that occur in a workplace context, using two broad categories: user type and meeting type.

At a minimum, two classes of potential users of a meeting information system can be identified. The first one, called in the following *group members*, includes people related to each other in the workplace, who often participate jointly in meetings (e.g. a design team), and people who, although might not regularly attend such meetings, have an interest in the group activities (e.g. a high-level manager monitoring the yearly progress of a specific team) [23, 18, 42, 11]. Briefly speaking, the users in this class have information needs related to information loss ("the failure to record important information, decisions and actions, and how this affects future

user	meeting type	use case	user needs
group member	local pre-recorded	1. previously attended meeting(s) search	- double-check action points for personal work - revisit tech details not clear in personal notes - summarize previous meeting(s) to prepare next agenda
		2. non-attended meeting(s) audit	- monitor project progress - examine the reasons for specific decisions - verify group cohesion / manager leadership
		3. on-line prior meeting(s) reminder	- revisit last-meeting agreements - resolve conflicts from last-minute meeting - follow up on unfinished issues
	local live	4. on-line meeting latecomer catch-up	- summarize meeting-in-progress - get action points for personal work - playback of critical issues
	remote live	5. enhanced meeting attendance (no video)	- generate description of participants' attitude - inform about identity of current conversants - inform about side conversations
		6. multi-task meeting monitor	- generate alerts for topic of interest - inform about heated discussions / action points - generate alert for personal presentation turn
external observer	local pre-recorded	7. semi-automatic group behavior annotator	- get speaker-turns as units for manual high-level annotation or statistical analysis - enrich manual annotations with automatic ones - produce co-occurrence stats for annotated behaviours
		8. social psychology teaching tool	- keyword-based search of segments with textbook behaviours - replay main excerpts of crisis management meetings - visualize long-term patterns over individuals and teams

Table 1: Meeting users, meeting types, use cases, and user needs (partly adapted from [52]) .

actions" [42]), and information mining (extracting trends from sets of meetings recorded over a possibly large period of time). The second class of potential users comprises a number of professional *external observers*, specialized in the study of group behavior, who are interested in defining, annotating, and detecting specific behaviors and trends from meeting sets, and thus might use a meeting information system as a work tool. This class of users includes social and organizational psychologists, instructors in these disciplines, and human resource officers among others [6, 25].

The meeting type gives rise to specific needs. The categorization adopted in this paper is based on the meeting's *physical location* and *time-of-recording*. The physical location can be local, where all participants are collocated (face-to-face), or remote, where some participants might attend from a separate location. Based on the time-of-recording, meetings can be pre-recorded, with a set of meetings available in a repository, or live, where meetings occur on-the-fly.

Various user/meeting pairs produce potential use cases for meeting access (browsing and retrieval) systems. In Table 1, we have listed eight of them (the first six are summarized from [52]), which assume that group members meet during a series of meetings. As can be seen from the Table, use cases have a degree of overlap (e.g. cases 1 and 3, 2 and 4, and 2 and 7). To further focus the discussion, in this paper we limit to review recent work on user requirement for local, pre-recorded meeting collections [23, 18, 42, 11]. Remote meetings are inserted in the large teleconferencing and computer-supported collaborative work (CSCW) domains, for which work on many aspects of user requirements exists [14, 36] but not discussed here due to lack of space. Finally, we are not aware of any comprehensive user requirements studies when users are professional external observers, probably due to the recent emergence of access systems to multimedia meeting collections.

Regarding local, pre-recorded meetings, various issues have been analyzed. The work in [18] questioned, among others, two aspects: the type of media items currently used to review meeting contents, and the reasons why people would use audio-visual recordings. It was found that public documents (including minutes and agendas) and personal notes (handwritten or electronic) are in major use, with audio-visual recordings being much less popular, arguably due to a lack of such resources [18]. It was also found that people would be most interested in accessing multimedia recordings to (1) keep accurate records, (2) understand unclear segments, (3) reexamine specific sections, (4) remember key people's statements, (5) recall ideas not stored in public or personal records, and (6) verify cases in which memory and written records are inconsistent [18].

The work in [42] confirmed some of these findings, with an interesting distinction between public and personal records. Binding in nature, public records -written minutes- are mostly useful to track group progress, to remember obligations, and to solve conflicts regarding obligations. At the same time, they sometimes lack accuracy, detail, context, and take effort to generate [42]. On the other hand, personal records are mainly used as reminders, as context providers for future actions, as minute backups, and as summaries to inform others. However, they can also be inaccurate, often cryptic (especially for others), and their generation limits participation in the actual meeting [42]. The study also higlighted the importance of looking at meetings not only from the single-meeting view but also from the collection perspective.

The approach taken in [23] used people placed in four scenarios -missed meeting, new employee, manager tracking project progress, and manager tracking employee performance- and generated queries for a hypothetical meeting retrieval system. An initial analysis of the queries categorized them into two broad classes, namely queries related to participants' interaction (including agreement/disagreement, acceptance/rejection, proposals, decisions, discussions, etc.), and queries related to the more general meeting domain (including dates/times, documents, participants, presentations, projects, tasks, topics, etc.) [23]. A more detailed analysis highlighted that that (1) queries often belong to both categories; (2) more queries belong to the second class; (3) a large number of queries involve only simple data processing to be answered satisfactorily; (4) some queries are about absent items (not present in the meeting); (5) audio and video are required to answer some of the queries (e.g. non-verbal ones); and (6) documents used or produced during meetings were often the queries' subject and are thus required in the collection [23]. Finally, the work in [11] seems to confirm some of the findings regarding media usage from [18, 42], and some of the preferred search styles from [23].

The discussed user requirements point to several relevant -and interrelated- research areas, defined in the next section.

3 Research areas in meetings

We now summarize what we consider to be basic research areas in the field, providing pointers to recent literature. Needless to say, these areas are rapidly advancing, and are not exclusive of the meeting domain.

1. Speech processing and analysis. What is said (and how it is said) is the first fundamental issue. However, speech in natural meetings is spontaneous and multi-party, containing disfluencies, no clear sentence boundaries, and significant overlapping, phenomena that constitute challenges for speech processing [34], from automatic transcription (see [53] for the most recent NIST automatic speech recognition (ASR) evaluation on meeting data) to higher-level tasks, like segmentation and classification of dialog acts (units that include backchannels, floor grabbers, questions, and statements) [2, 20].

2. Summary generation. Depending on the use case, summaries can vary both in form and in content, from an extractive textual summary to a selective replay of video segments with particularly interesting parts. Due to their conversational nature, speech in meetings often have low information content (compared to text documents), and many speech utterances relate to communication issues rather than to topics [7]. Various techniques have been adapted from text summarization in [45, 7, 29]. Other approaches conceptually related to summaries are those which attempt to detect meeting parts where participants are particularly engaged (called "hot-spots" in [43]). The relation between prosodic cues, dialog acts, and human-annotated hot-spots has been investigated, using speech utterances as basic units [43, 44]. We recently addressed a related task, namely the recognition of segments of high group interest-level from low-level audio-visual features [15], discussed in more detail in Section 4.

3. Document analysis. Text documents, including personal notes, slides, and e-documents, play an essential role in meetings. In addition to traditional text IR techniques (e.g. [39]), analyzing documents jointly with other media can be used for information verification and disambiguation, matching personal notes and audio-visual records, aligning references made in speech to documents, and creating richer text models combining text from written documents and speech for other tasks [26, 30].

4. Context modeling. To reexamine and understand information about meeting key phases (e.g. discussions that led to specific decisions), context could be extracted both from text content (personal notes) and from information other than spoken words (audio and video). Context can take a number of forms, including location [16], visual focus [37], addressee information [21], manifestations of emotional engagement like emphasis [22], and display of social signals like interest [15] and dominance [32]. The main challenge in all cases is modeling spontaneous natural behaviour.

5. Group interaction modeling. Meetings are a particular case of group interaction. Analyzing such interaction can provide, as stated above, important contextual information to enrich text or speech information. However, modeling meetings is relevant on its own for the external observer users defined in Section 2, where understanding human communication processes, both at short temporal scale (e.g. turn exchange dynamics) and long-term (e.g. influence and social connectivity) are key issues. We review our work on this area in section 4.

6. Long-term analysis. Analysis over long periods of time and several meetings is fundamental for project progress tracking, and discovery of group activities (e.g. usage of physical resources) and high-level trends (e.g. group cohesion). This is a very important area which, although appears in user requirement studies [42], has not been investigated much (exceptions are e.g. [31, 46]).

7. Media access. Given the rich and potentially large amount of available information, adequate ways of interacting with media to browse and retrieve information from meetings are needed. A recent review discussing existing systems to access multimedia meetings is [38].

Resources. Part of the research summarized above has been conducted using a number of multimedia meeting collections, each of which varies with respect to the sensor setup, the type of recorded meetings, the collection structure, and the type of existing annotations. Existing corpora include the ones by ISL (audio-only) [8], ICSI (audio-only) [19], with a dialog act annotation extension [33], NIST (audio-visual) [35], M4 (audio-visual) [24], AMI (audio, video, slides, whiteboard and handwritten notes) [9], and VACE (audio, video, and motion) [10]. These collections are at different stages of annotation and availability to the research community.

4 Group interaction modeling

In this section, we briefly describe our work on modeling of group interest-level, group activities, and influence. More details can be found in [15, 47, 48], respectively.

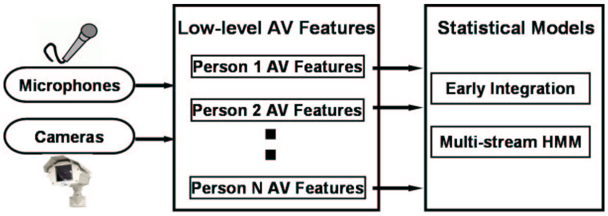

Figure 1: Detecting group interest-level in meetings using low-level features and HMMs.

case	$\alpha = 0$		$\alpha = 1$	
	pr	rc	pr	rc
1	0.54	0.85	0.70	0.34
2	0.54	0.85	0.73	0.42
3	0.59	0.84	0.75	0.55

Table 2: Precision/recall (pr/rc) values for two cases (higher is better). $\alpha = 0$ indicates that the system is trained by cross-validation to maximize recall; $\alpha = 1$ indicates that the system is trained to maximize precision. Cases 1-3 are described in the main text.

4.1 Modeling Group Interest-Level

As discussed before, finding relevant segments in meetings is important for summarization, browsing, and retrieval purposes. In [15], we defined relevance as the interest-level that meeting participants manifest as a group during the course of their interaction, and investigated the automatic recognition of segments of high-interest from audio-visual cues.

We addressed the problem using low-level audio-visual features and statistical models (Figure 1), with the goal of deriving, simultaneously, a segmentation for a meeting and the binary classification of its segments as having high or neutral interest-level. For sequence models, we investigated two classic Hidden Markov Model (HMM) recognition strategies. The first one is the basic early integration approach, where all desired streams (audio, visual, or audio-visual) are synchronized and concatenated to form the input observation vector. The second model is a multi-stream HMM (MS-HMM), which was only used for audio-visual fusion [12]. In this model, the audio and visual streams are trained independently, and the outputs of both modalities are merged at the state level during decoding, by a convex combination of the outputs, defined by a weight parameter.

The fully supervised approach called for human annotation of group interest-level for training (and testing) purposes. Such task required (1) multiple annotators, given that the task is to some degree subjective; (2) a criterion to evaluate whether there was reasonable agreement across annotators, both to define whether the task was computable and to set empirical performance bounds based on human performance; and (3) a mechanism to merge the multiple annotator judgements into a single annotation. The annotation was carried out on the M4 corpus [24], composed of 60 five-minute, four-participant meetings, which was recorded with three video cameras, a small circular 8-microphone array, and lapel microphones for each participant.

We extracted a set of audio-visual features, including audio features derived from microphone arrays and lapel microphones, and visual features extracted from skin color blobs from each participant. This initial audio-visual feature set was later used in an empirical feature selection procedure. We investigated various combination of models and features (i.e. audio-only, video-only, audio-video), and feature fusion at the group level.

We used the Expected Performance Curve (EPC) [4], based on precision/recall, to measure the performance of the models at the frame-level. As an illustration, the results obtained with three of the studied cases: (1) HMM, audio-only, individual features; (2) MS-HMM, audio-video, individual features; and (3) MS-HMM, audio-video, group features, are summarized in Table 2 (for complete results, see [15]). The analysis of the full results suggest that the audio modality is dominant for the task, that audio-visual fusion can improve performance, that MS-HMM with optimal combination weights outperforms early integration, and that feature

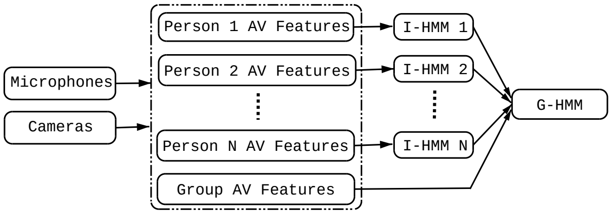

Figure 2: Multi-layer HMM for modeling group interaction.

fusion at the group level is beneficial.

4.2 Modeling Group Interaction with Layers

Viewed as a whole, a group in a meeting shares information, engages in discussions, and makes decisions, proceeding through diverse communication phases both in single meetings and during the course of long-term collaborative work. In [47], we attempted to structure meetings into sequences of high-level items (dubbed multimodal speaker turns), using a multi-layer HMM framework (Figure 2). We defined two sets of actions: group actions, which belong to the whole set of participants, such as *discussion* and *presentation*, and individual actions, belonging to specific persons, such as *writing* and *speaking*. Let I-HMM denote the lower recognition layer (individual action), and G-HMM denote the upper layer (group action). Each I-HMM receives as input audio-visual features extracted from each participant, and outputs posterior probabilities of the individual actions given the current observations. In turn, the G-HMM receives as input the output from several I-HMM (one per meeting participant), and a set of group features, directly extracted from the raw streams, which are not associated to any particular individual.

Compared with single-layer HMMs, multi-layer HMMs have the following advantages: (1) a single-layer HMM is defined on a possibly large observation space, which might face the problem of over-fitting with limited training data. In contrast, the layers in our approach are defined over small-dimensional observation spaces, resulting in more stable performance in cases of limited amount of training data. (2) The I-HMMs are person-independent, and in practice can be trained with much more data from different persons, thus better generalization performance can be expected. (3) The G-HMMs are less sensitive to variations in the low-level features because their observations are posterior-based features. (4) The two layers are trained independently, so we can explore different HMM combination systems. For example, we can replace the baseline I-HMMs with multi-stream HMMs. The framework thus becomes simpler to understand, and amenable to improvements at each layer.

Table 3 reports the performance in terms of action error rate (AER), equivalent to the word error rate in continuous ASR, for both multi-layer HMM and the single-layer HMM methods, tested on the M4 corpus. In this case, annotation for training and testing was available at the group level, given that the corpus was produced using scripts that defined meetings as sequences of the group actions we intended to recognize (the specific behavior of people was otherwise natural). Annotation at the individual action level was done by hand. Several configurations were compared, including audio-only, visual-only, early integration, multi-stream [12] and asynchronous HMMs [3]. Overall, the results suggest three main findings. First, the multi-layer HMM approach outperforms the single-layer one. Second, the use of AV features always outperforms the use of single modalities for both single-layer and multi-layer HMM, supporting the hypothesis that the group actions we defined are inherently multimodal. Third, the best I-HMM model is the asynchronous HMM (a model that explicitly accounts for variations of alignment between two data streams), which suggests that some asynchrony exists for the defined group actions, and that such asynchrony is reasonably captured by the model. A recent comparison of the layered HMM and other models on the same task appears in [1].

Method		AER (%)
Single-layer HMM	Visual only	48.20
	Audio only	36.70
	Early Integration	23.74
	Multi-stream	23.13
	Asynchronous	22.20
Multi-layer HMM	Visual only	42.45
	Audio only	32.37
	Early Integration	16.55
	Multi-stream	15.83
	Asynchronous	15.11

Table 3: Action error rates (AER) for single-layer and multi-layer HMM (lower is better).

Method	**KL divergence**
Random Guess	0.863
Speaking length	0.226
Influence model + Audio	0.135
Influence model + Language	0.106

Table 4: Influence modeling in meetings. Average KL divergence, computed between influence distributions estimated by human annotators and by automatic approaches (lower is better).

4.3 Modeling Influence

During the course of meetings, some people seem particularly capable of driving the conversation and dominating its outcome. These people, skilled at establishing the leadership, have the largest influence on a meeting, and often shift its focus when they speak. Can we tell who the most influential participant is? Can we quantify this amount of influence? How does the behavior of each individual affect the group decision-making? A computational model that addresses these questions involves challenges for the following reasons:

1. To build a model that can determine influence among meeting participants, we need to extract relevant features, with the assumption that influence can indeed be inferred from a set of low-level observations. In this sense, a large range of audio, visual and language features could be used. How to determine the most discriminative features is, however, a non-trivial task.

2. The task might be hard to evaluate. The manual annotation of influence of meeting participants is to some degree a subjective task, as a definite ground-truth does not exist.

3. To model a significant number of interacting people, the model requires an exponential number of parameters in the number of persons, which might make learning and inference intractable. This motivates the development of simplified models that at the same time retain representation power.

We have recently proposed a two-level influence model [48], which is a dynamic Bayesian network (DBN) with a two-level structure: the player level and the team level. The player level represents the actions of individual players, evolving based on their own Markovian dynamics. The team level represents group-level actions (the action belongs to the team as a whole, not to a particular player). The team state at the current timestep influences the players' states at the next timestep. In turn, the team state at the current timestep is also influenced by all the players' states at the current timestep.

For this task, the M4 corpus was once again annotated by hand using multiple annotators. In this case, the annotators were asked to define the distribution of influence over participants for each meeting in the corpus. The judgements coming from the different annotators were merged, after observing that there was sufficient agreement among them.

Regarding features, we extracted SRP-PHAT audio features to detect speaking turns in meetings [24]. Additionally, language features were extracted from manual speech transcripts. We compared our model with a method based on the speaking length (the proportion of time during which each participant speaks), and a method based on random guessing. To evaluate the results, we use the Kullback-Leibler (KL) divergence between the human-generated influence distributions and the automatically estimated distributions. The results are summarized in Table 4. On one hand, the results of the three methods: model+language, model+audio, and speaking-length are significantly better than the random result. On the other hand, using language features with our model produced the best performance. Importantly, our model (using either audio or language features) outperforms the speaking-length based method, which suggests that the learned influence distributions with our approach are in better accordance with the influence distributions from human judgement.

4.4 Open Issues

We conclude this section by discussing issues regarding computational models for human interaction analysis. While several solutions have already been proposed (as shown in the previous subsections) for modeling human interactions in the context of meeting related tasks, there are still several open problems in terms of machine learning.

One such problem is the lack of large and properly labeled meeting data sets. Indeed, most state-of-the-art techniques in meeting analysis assume that one has access to a large corpora of *training* meetings which are properly annotated according to the task. Hence, if the task is to identify high interest-level, one needs a collection annotated with such labels. Thus, for each new task, a different set of annotations is needed, with all the underlying human costs associated to it. Furthermore, in some cases, the annotation task in itself can be very difficult and noisy, giving rise to large variability among human annotators for the same data. For all these reasons, being able to estimate generic models based on raw data only, without any annotation, is very valuable, as large corpora of such data are much easier to obtain. These models could then be refined using some form of adaptation techniques (such as the Bayesian MAP adaptation [17]) on small but annotated training sets. More research is certainly needed in this direction.

Given the nature of meetings, which involves interactions among individuals, most current models start by extracting features from each individual present in the meeting, and then try to model their interaction. On the other hand, meetings often involve a varying number of participants including cases with individuals going in and out of the room during the same meeting. This poses a challenge at the early stages of modeling. A partial solution to this problem could come through the layered approach discussed in Section 4, where the first layer uses the same trained HMM for all individuals (and is thus independent of the number of individuals) in order to estimate individual activities, and then the second layer tries to combine the individual actions into group activities. This second step could be designed to integrate a variable number of individual high-level data.

From a more abstract modeling level, other problems are still open, and several of them are discussed in [5]. For instance, assuming each individual behavior is represented by a separate stream, and a single *group* model is used to incorporate all these streams, current modeling techniques, based on Markovian assumptions, often need exponential resources with respect to the number of streams, which quickly becomes intractable. Additionally, it is well known that long-term temporal dependencies are difficult to model without appropriate structural knowledge built in the model. This is still to be proposed in the context of human interactions. Finally, given the complex nature of human interactions, it should be important to be able to incorporate constraints (in the form of *prior* knowledge) at several levels of description (from the pixel level of the images, to the person-level, up to the group actions they overall performed).

5 Conclusions

We have presented a concise overview of some of the many facets of research on automatic extraction of information from multimedia meeting collections. Our intention was to provide the reader with pointers to recent literature on a number of tasks that, in our opinion, attempt (at least conceptually) to address the requirements of current and potential users of meeting information systems, with various degrees of robustness and direct applicability. In particular, we reviewed our work on modeling three aspects of group interaction. Overall, the meeting domain is still emerging, judging by the amount of work that has appeared recently, and by the challenges that remain unsolved. In our view, although it is likely that many of the existing analysis approaches will improve in the future, it will also be important to ground the discussion about technology progress on current and future user needs, if such technologies have any serious potential of becoming part of a real-world multimedia information system.

6 Acknowledgements

This work was supported by the Swiss National Center of Competence in Research on Interactive Multimodal Information Management (IM2), and the EC projects AMI (Augmented Multi-Party Interaction, pub. AMI-126) and PASCAL (Pattern Analysis, Statistical Modeling, and Computational Learning) through the Swiss OFES. We thank several of our colleagues from the M4 and AMI projects for discussions and interaction over time, which have contributed to shape many of the views expressed here.

References

[1] M. Al-Hames, A. Dielmann, D. Gatica-Perez, S. Reiter, S. Renals, G. Rigoll, and D. Zhang, "Multimodal Integration for Meeting Group Action Segmentation and Recognition," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh, Jul. 2005.

[2] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, Mar. 2005.

[3] S. Bengio, "An asynchronous Hidden Markov Model for audio-visual speech recognition," in *Proc. Advances in Neural Information Processing Systems, (NIPS 15)*, Vancouver, Dec. 2002.

[4] S. Bengio and J. Mariethoz, "The expected performance curve: a new assessment measure for person authentication," in *Proc. Odyssey*, Toledo, May 2004.

[5] S. Bengio and H. Bourlard, "Multi channel sequence processing," in *Proc. PASCAL Machine Learning Workshop*, Sheffield, Sep. 2004.

[6] R.F. Bales, *Interaction Process Analysis: a method for the study of small groups*, Addison-Wesley, 1951.

[7] A. H. Buist, W. Kraaij, and S. Raaijmakers, "Automatic summarization of meeting data: A feasibility study," in *Proc. Meeting of Computational Linguistics in the Netherlands (CLIN)*, Leiden, Dec. 2004.

[8] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech style," in *Proc. ICSLP*, Denver, Sep. 2002.

[9] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh, Jul. 2005.

[10] L. Chen, R. Travis Rose, F. Parrill, X. Han, J. Tu, Z. Huang, M. Harper, F. Quek, D. McNeill, R. Tuttle, and T. Huang, "VACE multimodal meeting corpus," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh, Jul. 2005.

[11] A. Cremers and B. Hilhorst, "What was discussed by whom, how, when and where? Personalized browsing of annotated multimedia meeting recordings," in *Proc. Int. Conf. on Human-Computer Interaction (HCI International)*, Las Vegas, Jul. 2005.

[12] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. on Multimedia*, vol. 2, no. 3, pp. 141–151, Sep. 2000.

[13] B. Erol and Y. Li, “An overview of technologies for e-meeting and e-lecture,” in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME)*, Amsterdam, Jul. 2005.

[14] S. Elrod, R. Bruce, R. Gold, D. Goldberg, and F. Halasz, “LiveBoard: a large interactive display supporting group meetings, presentations and remote collaboration,” in *Proc. ACM Conf. on Human Factors in Computing Systems (CHI)*, Monterey, May 1992.

[15] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, “Detecting group interest-level in meetings,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, Mar. 2005.

[16] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, “Multimodal multispeaker probabilistic tracking in meetings,” in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, Trento, Oct. 2005.

[17] J.L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture obervation of Markov chains,” *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 290–298, 1994.

[18] A. Jaimes, K. Omura, T. Nagamine, and K. Hirata, “Memory cues for meeting video retrieval,” in *Proc. ACM Int. Conf. on Multimedia, Workshop on Continuous Archival and Retrieval of Personal Experiences (ACM MM-CARPE)*, New York, Oct. 2004.

[19] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, “The ICSI meeting corpus,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Hong-Kong, Apr. 2003.

[20] G. Ji and J. Bilmes, “Dialog act tagging using graphical models,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, Mar. 2005.

[21] N Jovanovic and R. op den Akker, “Towards automatic addressee identification in multi-party dialogues,” in *Proc. SIGDial Workshop on Discourse and Dialogue*, Boston, Apr. 2004.

[22] L. Kennedy and D. Ellis, “Pitch-based emphasis detection for characterization of meeting recordings,” in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Virgin Islands, Dec. 2003.

[23] A. Lisowska, A. Popescu-Belis, and S. Armstrong, “User query analysis for the specification and evaluation of a dialogue processing and retrieval system,” in *Proc. Int. Conf. on Language Resources and Evaluation (LREC)*, Lisbon, May 2004.

[24] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, “Automatic analysis of multimodal group actions in meetings,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 305–317, Mar. 2005.

[25] J.E. McGrath, *Groups: Interaction and Performance*, Prentice-Hall, 1984.

[26] D. Mekhaldi, D. Lalanne, and R. Ingold, “Thematic segmentation of meetings through document/speech alignment,” in *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, New York, Nov. 2004.

[27] T. P. Moran, S. Palen, L.and Harrison, P. Chiu, D. Kimber, S. L. Minneman, B. van Melle, and P. Zellweger, “I’ll get that off the audio: a case study of salvaging captured multimedia meeting records,” in *Proc. ACM Int. Conf. on Human Factors in Computing Systems (CHI)*, Atlanta, Mar. 1997.

[28] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, “The meeting project at ICSI,” in *Proc. Human Language Technology Conf. (HLT)*, San Diego, CA, March 2001.

[29] G. Murray, S. Renals, and J. Carletta, “Extractive summarization of meeting recordings,” in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, Lisbon, Sep. 2005.

[30] A. Popescu-Belis and D. Lalanne, "Detection and resolution of references to meeting documents," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh, Jul. 2005.

[31] S. Renals and D. Ellis, "Audio information access from meeting rooms," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, 2003.

[32] R. Rienks and D. Heylen, "Automatic dominance detection in meetings using support vector machines," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh, Jul. 2005.

[33] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. HLT-NAACL SIGDIAL Workshop*, Boston, Apr. 2004.

[34] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, Lisbon, Sep. 2005.

[35] V. Stanford, J. Garofolo, , and M. Michel, "The nist smart space and meeting room projects: Signals, acquisition, annotation, and metrics," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, 2003.

[36] N. Streitz, J. Geissler, J. Haake, and J. Hol, "DOLPHIN: integrated meeting support across local and remote desktop environments and LiveBoards," in *Proc. ACM Conf. on Computer Supported Cooperative Work (CSCW)*, Chapel Hill, Oct. 1994.

[37] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing based on multiple cues," *IEEE IEEE Trans. on Neural Networks*, vol. 13, no. 4, pp. 928–938, 2002.

[38] S. Tucker and S. Whittaker, "Accessing multimodal meeting data: Systems, problems and possibilities," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Martigny, Jun. 2004.

[39] A. Vinciarelli and J.-M. Odobez, "Application of information retrieval techniques to presentation slides," *IEEE Trans. on Multimedia*, 2005, in press.

[40] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner, "Advances in automatic meeting record creation and access," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, May 2001.

[41] S. Whittaker, P. Hyland, and M. Wiley, "Filochat: handwritten notes provide access to recorded conversations," in *Proc. ACM Int. Conf. on Human Factors in Computing Systems (CHI)*, Boston, Apr. 1994.

[42] S. Whittaker, R. Laban, and S. Tucker, "Analysing meeting records: an ethnographic study and technological implications," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh, Jul. 2005.

[43] B. Wrede and E. Shriberg, "Spotting hotspots in meetings: Human judgments and prosodic cues," in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, Geneva, Sep. 2003.

[44] B. Wrede and E. Shriberg, "The relationship between dialogue acts and hot spots in meetings," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Virgin Islands, Dec. 2003.

[45] K. Zechner, "Automatic summarization of open-domain multiparty dialogues in diverse genres.," *Computational Linguistics*, vol. 28, pp. 447–485, 2002.

[46] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud, "Multimodal group action clustering in meetings," in *Proc. ACM Int. Conf. on Multimedia, Workshop on Video Surveillance and Sensor Networks (ACM MM-VSSN)*, New York, Oct. 2004.

[47] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Modeling individual and group actions in meetings with layered HMMs," *IEEE Trans. on Multimedia*, 2005, in press.

[48] D. Zhang, D. Gatica-Perez, S. Bengio, and D. Roy, "Learning Influence among Interacting Markov Chains," in *Proc. Advances in Neural Information Processing Systems (NIPS 18)*, Vancouver, Dec. 2005.

[49] Augmented Multi-Party Interaction (AMI) project, www.amiproject.org.

[50] Interactive Multimodal Information Management (IM2) project, www.im2.ch.

[51] MultiModal Meeting Manager (M4) project, www.m4project.org.

[52] AMI project, "Use cases and user requirements," Public deliverable D6.2, Apr. 2005.

[53] NIST, *Proc. Rich Transcription 2005 Spring Meeting Recognition Evaluation Workshop*, Edinburgh, Jul. 2005.