Recording, Indexing, Summarizing, and Accessing Meeting Videos: An Overview of the AMI Project

Alejandro Jaimes[1], Herve Bourlard[1], Steve Renals[2], and Jean Carletta[2]
[1]*IDIAP Research Institute,* [2]*University of Edinburgh*
alex.jaimes@idiap.ch

Abstract

In this paper we give an overview of the AMI project. AMI developed the following: (1) an infrastructure for recording meetings using multiple microphones and cameras; (2) a one hundred hour, manually annotated meeting corpus; (3) a number of techniques for indexing, and summarizing of meeting videos using automatic speech recognition and computer vision, and (4) an extensible framework for browsing, and searching of meeting videos. We give an overview of the various techniques developed in AMI, their integration into our meeting browser framework, and future plans for AMIDA (Augmented Multiparty Interaction with Distant Access), the follow-up project to AMI.

1. Introduction

Over the last few years research interest in recording, archiving, and retrieving of meeting videos has increased significantly. This is due to major drops in hardware costs, broadband availability (for remote meetings), and concerns by corporations about record keeping (auditing decision-making, corporate memory, and complying with regulatory requirements, etc.).

Meetings play a crucial role in the generation of ideas, documents, relationships, and actions within an organization. The wealth of information exchanged in meetings, however, is often lost because manual creation of meeting minutes is subjective, incomplete, and captures only a fraction of the information. Audio-visual recording of meetings is therefore attractive, but leads to many practical challenges, from the infrastructure to record the meetings to the archival, indexing, and retrieval of relevant meeting segments. Given the number of meetings in most organizations, efficient and effective recording and access to meeting videos is of extreme importance, making research in content-based indexing and retrieval of meeting videos an important research area, not only because of its potential impact, but also because it requires combining research in several disciplines (e.g., speech recognition, computer vision, etc.).

In this paper, we describe the AMI project. AMI deals with meeting videos throughout the media production chain: from modeling of meetings, to recording infrastructure and recording, to multimodal, automatic indexing, retrieval, and browsing of meeting videos. We give a general overview of each of the components above and discuss use of AMI technologies within the framework we have developed for browsing, searching, and summarization of meeting videos. The goal of this paper and its main contribution, therefore, is to give an overview of the technologies developed in the project and their integration within applications for searching and browsing.

Related Work. Meeting room projects focus on portable recorders [10], speech [4], modeling [1], video capture [13], and others. The AMI project's components build on and improve the state-of-the art in many areas, and since this paper gives a general overview we refer interested readers to specific AMI publications [6] for details on how specific techniques developed within AMI differ from related work.

2. Instrumented Meeting Rooms & Ami Corpus

Three meeting rooms were designed and constructed at AMI partners IDIAP, TNO and the University of Edinburgh. These rooms, which were designed for the recording of videos of four person meetings, all contained a set of standardized recording equipment (plus additional cameras, microphone arrays, and binaural manikins):

- 6 cameras: 4 providing close-up views of the participants, 2 providing a view of the whole room;
- 12 microphones: a headset microphone per participant and an 8-element circular microphone array;
- data projector capture (VGA);
- white board capture and digital pen capture.

The meeting rooms were used to record the AMI Meeting Corpus [6], which consists of 100 hours of meeting recordings. The corpus includes manually produced orthographic transcriptions of the spoken dialogues, aligned at the word level with the common time line, and annotations describing participant behavior during the meetings (e.g., dialogue acts; topic segmentation; extractive and abstractive summaries; named entities; limited forms of head gesture, hand gesture and gaze direction; movement around the room; emotional state; head localization, etc.).

The corpus consists of two types of meetings: (1) remote control design scenario (approx. 2/3 of AMI corpus), (2) free topic. In the design scenario, each group of four participants had four meetings and given tasks to complete between meetings (with the final goal of designing a T.V. Remote control). Participant roles were driven in real-time by emails and web information. This control made it easier to understand the content of the meetings, enabled the construction of ontologies, and the building of outcome measures (e.g., preferred design output). The meetings are also replicable, enabling system-level evaluations. Free topic meetings were naturally occurring meetings in a range of domains. The project further developed NXT (NITE XML Toolkit [7]), an open source XML-based infrastructure for the annotation and management of multimodal recordings. NXT consists of libraries from which user interfaces for annotating and searching annotations of multi-modal data sets can be easily built. Within AMI, new tools for annotation were created, for instance for dialogue acts, named entities, topic segmentation, summarization, and a generic time-aligned coder and display.

3. Audio-Visual Processing

AMI work in audio-visual processing was primarily concerned with the development of algorithms that, given raw audio-visual streams, can automatically answer each of the following questions [1]:

- What has been said during the meeting? (Speech recognition)
- What acoustic events and keywords occur in the meeting? (Keyword spotting)
- Who and where are the persons in the meeting? (Localization and tracking)
- Who in the meeting is acting or speaking? (Speaker tracking)
- How do people act in the meeting? (Gesture and action recognition)
- What are the participants' emotions in the meeting? (Emotion)
- Where or what is the focus of attention in meetings? (Focus of attention)

Speech recognition. AMI developed systems for two types of microphone configurations in the instrumented meeting rooms (close-talking headset microphones and tabletop microphone arrays), focusing on the headset microphone conditions to develop core acoustic modeling approaches, but with an overall orientation to tabletop microphone arrays, which are less intrusive [15]. The AMI speech recognition effort addressed several research issues including the following:

- microphone array beamforming: filtering and combining the individual microphone signals to enhance signals coming from a particular location (and suppressing competing locations);
- development of novel acoustic parameterizations, including approaches based on posterior probability estimation;
- automatic construction of domain-specific language models using text extracted from the web;
- acoustic segmentation;
- development of a flexible large vocabulary decoder, based on a weighted finite state transducer formalism.

AMI developed an evaluation framework that is generic, flexible, comparable, and that allows us to conduct research and development in a stable environment. Using this framework, our system obtains exceptionally good results on AMI meeting data; in international technology evaluations organised by NIST, no other system was significantly more accurate than the AMI system on close-talking microphones [16]. This system has been used to decode the complete AMI corpus (using an n-fold cross-validation technique). The transcriptions have been used for tasks such as summarization and topic segmentation.

Keyword spotting. In acoustic keyword spotting (KWS), the goal is to find keywords and their position in speech data. AMI developed three approaches: acoustic, LVCSR, and a hybrid approach [17]. In the *acoustic approach*, a keyword score is obtained by comparing the posterior probability of the keyword phonetic model, with a background model. This is very fast since many of the key parameters may be precomputed. It is relatively precise (the precision increases with the length of the keyword) and any word can be searched provided its phonetic form is available. It is ideal for on-line applications (such as monitoring remote meetings), but it is not suitable for browsing huge archives, as all of the acoustic data must be processed for each search. The *LVCSR lattice approach* locates the keywords in lattices generated by

a large vocabulary continuous speech recognition system. Given the output of the speech recognizer, this approach is very fast, but it is accurate only for frequently occurring words. There is a degradation in performance for less common words, which is a drawback since these words (such as technical terms and proper names) carry most of the information and are likely to be searched by users. Therefore, this approach has to be complemented by a method unconstrained by the recognition vocabulary. The *hybrid phoneme lattice approach* is based on the construction of graphs of phoneme probabilities, from which the phonetic form of the keyword may be extracted. This is a reasonable compromise in terms of accuracy and speed. Currently, AMI work on indexing phoneme lattices using tri-phoneme sequences is advancing and preliminary results show good accuracy/speed trade-off for rare words.

Speaker tracking. The objective of speaker tracking is to segment, cluster and recognize the speakers in a meeting, based on their speech. The first approach developed in AMI uses the acoustic contents of the microphone signal to segment and cluster speakers. In the NIST evaluations this system produced very good results for speech activity detection (the lowest error rate reported) and for speaker diarization (who spoke when). The second approach developed in AMI, based on cross-correlations between microphone signals operates in real time, and has been integrated with the online keyword spotter [11] .

Localization and tracking. Location coordinates of each person in the meeting are an essential input to various meeting analysis tasks, including focus of attention and action recognition. The steps required are identification, localization, and tracking. For identification, generative approaches have proven to be the most robust so in AMI a variety of models with different trade offs between speed and accuracy have been used (e.g., based on Gaussian mixtures and HMMs). The algorithms have been developed as a machine vision package for the open source machine learning library, TORCH (extended within AMI (http://www.torch.ch)). For localization and tracking AMI developed, applied, and evaluated four different methods including approaches based on dynamic Bayesian networks, active shape trackers using particle filters, and face trackers based on skin colour.

Gesture and action recognition. We have defined a set of actions and gestures that are relevant for meetings (e.g., hand, body, and head gestures such as pointing, writing, standing up, or nodding). Special attention has been paid to negative signals, such as a negative response to a yes-no question, usually characterized by a head shake. This kind of gesture contains important information about the decision making in meetings, but can be very subtle and involve little head movement, making automatic detection very difficult. For gesture recognition two methods were applied: *Bayesian Information Criterion* and an *Activity Measure* approach. For each person in the meeting, the 2D location of the head and hands, a set of nine 3D joint locations, and a set of ten joint angles were extracted. In addition classification was performed in the segmented data. Due to the temporal character of gestures the focus was on different HMM methods. Gestures like standing up and important speech supporting gestures produced satisfactory results (100% and 85% recognition rate, respectively). However, the results for the detection of negative signals were not significantly better than guessing. Detecting gestures such as shaking or nodding is challenging and requires disambiguating the meaning of very subtle head movements.

Focus of Attention. Gaze detection requires higher resolution of facial images than what is available in the AMI corpus. As an approximation, we have developed algorithms for tracking the head and estimating its pose, based on a Bayesian filtering framework, which is then solved through sampling techniques. Results (evaluated on 8 minutes of meeting recordings involving a total of 8 people) were good, with a majority of head pan (resp. tilt) angular errors smaller than 10 (resp. 18) degrees. As expected, we found a variation of results among individuals, depending on their resemblance with people in the appearance training set. In addition, we formulated focus of attention (FoA) as a classification task by automatically classifying FoA into one of the following categories: meeting participants, objects in the meeting room, and an “unfocused” location. Experiments using the ground truth head-pose pointing vectors resulted in frame-based classification rate of 68% and 47%, depending on the person's position in the smart meeting room. Accuracy is lower than reported in other works, mainly because of the complexity of the scenes and number of categories. Exploiting other features/modalities (e.g speaking status) in addition to the head pose can be used to disambiguate FoA classification. We found that using the estimated head-pose instead of the ground truth did not degrade the results strongly (about 9% decrease, thus much less than the differences w.r.t. position in the meeting room), which was encouraging given the difficulty of the task. We also found that there was a large variation of recognition amongst individuals, which directly calls for approaches such as

Maximum A Posteriori techniques for the FoA recognition (the topic of current research).

4. Content Extraction

Dialogue acts are labels for utterances which roughly categorize the speaker's intention. They are useful, for example as part of a browser which highlights all points where a suggestion or offer was recognized. Dialogue acts also serve as elementary units, upon which further structuring or discourse processing may be based (e.g., summarization). The following dialog act labels were used:

- Information exchange: giving and eliciting information;
- Possible actions: making or eliciting suggestions or offers;
- Commenting on the discussion: making or eliciting assessments and comments about understanding;
- Social acts: expressing positive or negative feelings towards individuals or the group;
- Other: utterances which convey an intention, but do not fit into the four previous categories;
- Back channel, Stall and Fragment: utterances without content, which allow complete segmentation of the material.

We have used combinations of machine learning based on a multimodal set of features, including a word-based language model, prosodic features (based on duration, energy and intonation), context features (e.g., speaker overlap), and discourse features (history of previously recognized dialogue acts). Using generative models that explicitly take account of the dependence on multiple streams of data (such as dynamic Bayesian networks, factored language models, and hidden event language models) we have obtained state-of-the-art results for dialogue act segmentation. Interestingly, although the best approach to dialogue act segmentation involves jointly segmenting and labeling the dialogue act sequence, we have found that the labeling may be substantially improved by re tagging using discriminative approaches, in particular conditional random fields. Comparing the performance on automatically transcribed speech with human transcribed speech, we find that the performance of dialogue act recognition drops by about 10%.

Topic segmentation. The aim of topic segmentation is to automatically infer the sequential structure of the meeting by topic (and sub-topic); it differs from dialogue act recognition in that the fundamental units (topics) are typically many minutes in duration. We have explored two basic approaches to this task. An unsupervised approach, LCSeg automatically infers (without training) topic boundaries as points where the statistics of text change significantly. The supervised approach, on the other hand, learns topic boundaries based on a hand-annotated training set. An advantage of the supervised approach is that it is possible to use additional features relating to prosody (e.g., pauses) and the structure of the conversation (e.g., speaker overlap). These additional features are also relatively independent of errors in the automatic speech transcription. We have also developed approaches to automatically generate labels for topics, based on the statistics of the automatically transcribed words that make up a topic.

Summarization. We have investigated two distinct ways of constructing summaries of a meeting. Extractive techniques construct summaries by locating the most relevant parts of a meeting and concatenating them together to provide a 'cut-and-paste' summary, which may be textual or multimodal. Abstractive summaries, on the other hand, are similar to what a human summarizer might construct, generating new text to succinctly describe the meeting. Abstractive summarization is more challenging than extractive summarization, and requires relatively deep domain knowledge. Our approach to extractive summarization is based on automatically extracting relevant dialogue acts from a meeting. It thus requires (as a minimum) the automatic speech transcription and the dialogue act segmentation modules described above. Lexical information is clearly extremely important for this task, but we have found it beneficial to augment information derived from the transcription with speaker features (relating to activity, dominance and overlap), structural features (the length and position of dialogue acts), prosody, and discourse cues (phrases which signal likely relevance). All of these features are important to develop accurate methods for extractive summarization. We have also explored reduced dimension representations of text, based on latent semantic analysis, which also add precision to the summarization. Using an evaluation measure referred to as weighted precision, we have discovered that it is possible to reliably extract the most relevant dialogue acts, even in the presence of speech recognition errors. We have explored “dialogue act compression,” in which the extracted dialogue acts are condensed by removing irrelevant portions. Again, taking account of speech features such as the overall intonation contour of the dialogue act helps to improve the overall performance. We have also implemented a prototype abstractive summarization system, based on an ontology of the AMI scenario meetings, together with annotations of propositional content, and the topic structure of the meetings. Given these annotations an

ontological representation is built, which is then passed to a natural language generation component which produces a one paragraph summary of the meeting.

Influence and dominance detection. Person-to-group influence (i.e., influence of a person over the group) is estimated from audio features with a framework based on a two-level Dynamic Bayesian Network, in which an influence distribution is defined as the prior probability of individual state streams contributing to the group state stream. Such a distribution can be automatically estimated from data and was tested on AMI spoke data. Dominance relations between meeting participants has also been inferred. Using SVMs we were able to predict who is more, less or normally dominant in a meeting with an accuracy of 75%.

Video content extraction. We have developed "automatic camera operator" algorithms based on extracted video and audio features to perform this operation. Subjective evaluation with users indicated that the deployed algorithms were functionally acceptable, but were of significantly lower aesthetic quality compared with human production. We have also developed methods for identifying "hot-spots" such as laughter, directly from video features based on things such as motion and texture.

5. AMI Meeting Browsers

Many AMI technologies are demonstrated within a Java-based browsing framework, referred to as JFerret. JFerret is a multimedia browser that is extremely flexible, enabling almost any user interface to be composed, using a combination of plug-in modules. An XML configuration specifies which plug-in components to use, how to arrange them visually, and how they communicate with each other. JFerret comes with a library of pre-defined plugins, for presentation of video, audio, slides, annotation time-lines, controls, and so on, and it is straightforward to write new plugins. This has been the main route to demonstration for many of the technologies described in previous sections. Java allows the application to run cross-platform, either as a applet (inside a web-browser) or as a stand-alone application. An example JFerret configuration, enables browsing via keyword search on the speech-recognized transcript, search within captured slides, and browsing by speaker activity. Time-synchronized recordings that may be browsed include multiple video and audio streams and white board capture. Other semantically rich browser components that have been constructed include direct keyword-spotting, video hot spots, and argumentation.

We have also begun to explore techniques for time-based media compression, since this can clearly contribute to efficient browsing of recorded meetings. Time-based compression can be done in three major ways: 1) speech speedup, 2) excision of less important parts, and 3) simultaneous presentation of speech from two locations. Two interactive prototypes for accelerated listening of recorded speech have been implemented. One prototype provides support for speed controls as well as skipping ahead and back based on speaker segmentations. The other prototype presents two parts of the meeting simultaneously using binaural in two different locations so that the user can listen to one part of the meeting while monitoring another part. We also devised a PDA-based wireless presentation system, including recording of slide presentations, which was integrated with the meeting browser using VNC.

Evaluation. AMI scientists have been closely involved in several international evaluation efforts such as the NIST Meeting Recognition evaluation of speech recognition and speaker diarization in meetings, for which the AMI corpus has been one of the main data sources. AMI has also participated in the CLEAR evaluations of focus of attention and face detection. Additionally, the AMI corpus, together with speech recognition output, has been provided to the Cross Language Evaluation Forum (CLEF) for their 2007 evaluation on cross-lingual question answering. In addition, AMI developed a framework for extrinsic evaluation of browser components, called the Browser Evaluation Test (BET). The BET provides a framework for the comparison of arbitrary meeting browser setups, where setups differ in terms of which content extraction or abstraction components are employed. The BET consists of a set of experiments in which test subjects have to answer true/false questions about observations of interest for a meeting recording. The test subject uses the browser under test to answer these questions, given a time limit (typically half the meeting length). This framework has proven to be a successful way to evaluate browser components.

We have also developed a task-based evaluation that is supported by the design of the AMI corpus (about 70% of corpus meetings are based on a replicable design team scenario). In the task-based evaluation, a new team takes over for the fourth meeting, with access to the previous three meetings. The evaluation compares team performance in the existing case with basic meeting records (including powerpoint files, emails and minutes), with a basic AMI meeting browser, and with a task-based browser. The task-based evaluation is in terms of both objective measures

such as design quality, meeting duration, assessment of outcome, and behaviourial measures of leadership, and subjective measures including browser usability, workload (mental effort), and group process.

6. Conclusions and Future Work

We have provided an overview of the AMI project. The major achievements of AMI are in six areas: Instrumented meeting rooms (development of a recording infrastructure, based on instrumentation of meeting rooms, in which we can capture all aspects of interaction in a meeting, in a time synchronized manner), the AMI Corpus (a 100 hour corpus of recorded meetings, with multiple time synchronized signals across several modalities, annotated at many different levels), audio-video processing (significant advances in several areas including speech recognition, audio-video localization and tracking, and detection of focus of attention), content extraction (new state-of-the-art techniques in several areas such as summarization and dialogue act recognition), integrated demonstrations (AMI developed an integrated browsing framework in which the outputs of multimodal recognition and content extraction modules may be incorporated as plugins or data streams), and evaluation (novel frameworks for system evaluation). For each of the areas described there are many ongoing improvements and plans for future work. In general, improving robustness, speed, and accuracy are important issues, as well as scaling the techniques to deal with larger amounts of data. Within the new AMIDA project [6] we are working on improving many of the techniques, paying particular attention to their integration into a framework of "meeting assistants" that can perform in close-to real-time (i.e., delays of several seconds or even minutes may be acceptable) within applications that integrate these techniques for use during, and between meetings, in remote and co-located settings.

Acknowledgements. This work has been performed by the AMI consortium, which is a 6th Framework Research Programme of the European Union (EU), contract number: IST-033812. The authors would like to thank the EU for the financial support and the partners within the consortium for a fruitful collaboration. Part of the work presented here was also funded the Swiss National Science Foundation, through the National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)", http://www.im2.ch. Special thanks to John Dines for useful comments.

10. References

[1] M. Al-Hames, et. al., "Audio-Visual Processing in Meetings: Seven Questions and Current AMI Answers," in Wksp on Machine Learning for Multimodal Interaction (MLMI), Washington DC, May 2006.

[2] C. Costa, P. Antunes, and J. Dias, "A Model for Organizational Integration of Meeting Outcomes," in Contemporary Trends in Systems Development, M.K. Sein et. al., eds, Kluwer Plenum, 2001.

[3] R. Cutler, et. al., "Distributed Meetings: A Meeting Capture and Broadcasting System," Proc. ACM Multimedia 2002, Juan Les Pines, France, 2002.

[4] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg, "Meeting Recorder Project: Dialog Act Labeling Guide," ICSI Technical Report TR-04-002, 2004.

[5] W. Geyer, H. Ritcher, and G. Abowd, "Making Multimedia Meeting Records More Meaningful," IEEE ICME 2003, Baltimore, MD, July 2003.

[6] www.amiproject.org; http://corpus.amiproject.org

[7] http://sourceforge.net/projects/nite/

[8] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection Of Agreement vs. Disagreement In Meetings: Training With Unlabeled Data," Proc. HLT-NAACL Conference, Edmonton, Canada, May 2003

[9] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, B. Wrede. "The ICSI Meeting Project: Resources and Research," NIST ICASSP 2004 Meeting Recognition Workshop, Montreal, May 2004.

[10] D.-S. Lee, B. Erol, J. Graham, H.J. Hull, and N. Murata, "Portable Meeting Recorder," in proc. ACM Multimedia 2002, Juan Les Pines, France, 2002.

[11] P. Motlicek, L. Burget, and J. Cernocky. Non-parametric speaker turn segmentation of meeting data. In Proceedings Eurospeech, 2005.

[12] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling human interaction in meetings. In IEEE ICASSP, 2003.

[13] Y. Rui, A. Gupta, J. Grudin and L. He, "Automating lecture capture and broadcast: technology and videography," ACM Multimedia Systems Journal, 3-15, Springer V., 2004.

[14] B. Wrede and E. Shriberg, "Spotting Hot Spots in Meetings: Human Judgements and Prosodic Cues," in EUROSPEECH 2003, Geneva, September 2003.

[15] T. Hain, L. Burget, J. Dines, G. Garau, V. Wan, M. Karafiat, J. Vepa and M. Lincoln, "The AMI system for the transcription of speech in meetings", Proc. ICASSP Honolulu, Hawaii, 2007.

[16] S. Renals,, S. Bengio, J. Fiskus (Eds). Machine Learning for Multimodal Interaction: 3rd Intl. Workshop, MLMI'06. Springer Lecture Notes in Computer Science, Bethesda, MD, USA, 2006,

[17] I. Szöke, P. Schwarz, L. Burget, M. Fapso, M. Karafiát, J. Cernocký, P. Matejka. "Comparison of Keyword Spotting Approaches for Informal Continuous Speech", Proc. Interspeech '05 (Eurospeech), Lisabon, Portugal, 2005.