Audio-Visual Processing in Meetings: Seven Questions and Current AMI Answers

Marc Al-Hames[1], Thomas Hain[2], Jan Cernocky[3], Sascha Schreiber[1], Mannes Poel[4], Ronald Müller[1], Sebastien Marcel[5], David van Leeuwen[6], Jean-Marc Odobez[5], Sileye Ba[5], Herve Bourlard[5], Fabien Cardinaux[5], Daniel Gatica-Perez[5], Adam Janin[8], Petr Motlicek[3,5], Stephan Reiter[1], Steve Renals[7], Jeroen van Rest[6], Rutger Rienks[4], Gerhard Rigoll[1], Kevin Smith[5], Andrew Thean[6], and Pavel Zemcik[3] **

[1] Institute for Human-Machine-Communication, Technische Universität München
[2] Department of Computer Science, University of Sheffield
[3] Faculty of Information Technology, Brno University of Technology
[4] Department of Computer Science, University of Twente
[5] IDIAP Research Institute and Ecole Polytechnique Federale de Lausanne (EPFL)
[6] Netherlands Organisation for Applied Scientific Research (TNO)
[7] Centre for Speech Technology Research, University of Edinburgh
[8] International Computer Science Institute, Berkeley, CA

Abstract. The project Augmented Multi-party Interaction (AMI) is concerned with the development of meeting browsers and remote meeting assistants for instrumented meeting rooms – and the required component technologies R&D themes: group dynamics, audio, visual, and multimodal processing, content abstraction, and human-computer interaction. The audio-visual processing workpackage within AMI addresses the automatic recognition from audio, video, and combined audio-video streams, that have been recorded during meetings. In this article we describe the progress that has been made in the first two years of the project. We show how the large problem of audio-visual processing in meetings can be split into seven questions, like "Who is acting during the meeting?". We then show which algorithms and methods have been developed and evaluated for the automatic answering of these questions.

1 Introduction

Large parts of our working days are consumed by meetings and conferences. Unfortunately a lot of them are neither efficient, nor especially successful. In a recent study [12] people were asked to select emotion terms that they thought would be frequently perceived in a meeting. The top answer – mentioned from more than two third of the participants – was "boring"; furthermore nearly one third mentioned "annoyed" as a frequently perceived emotion. This implies that many people feel meetings are nothing else, but flogging a dead horse.

** This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811).

Things get from bad to worse if transcriptions are required to recapitulate decisions or to share information with people who have not attended the meeting. There are different types of meeting transcriptions: they can either be written by a person involved in the meeting and are therefore often not exhaustive and usually from the particular perspective of this person. Sometimes they are only hand-written drafts that can not easily be shared. The second type are professional minutes, written by a person especially chosen to minute the meeting, usually not involved in the meeting. They require a lot of effort, but are usually detailed and can be shared (if somebody indeed takes the time to read over them). The third and most common transcript is no transcript at all.

Projects, like the ICSI meeting project [14], Computers in the human interaction loop (CHIL) [29], or Augmented Multi-party Interaction (AMI) [7] try to overcome these drawbacks of meetings, lectures, and conferences. They deal with the automatic transcription, analysis, and summarisation of multi-party interactions and aim to both improve the efficiency, as well as to allow a later recapitulation of the meeting content, e.g with a meeting browser [30]. The project AMI is especially concerned with the development of meeting browsers and remote meeting assistants for instrumented meeting rooms – and the required component technologies R&D themes: group dynamics, audio, visual, and multimodal processing, content abstraction, and human-computer interaction."Smart meeting rooms" are equipped with audio-visual recording equipment and a huge range of data is captured during the meetings. A corpus of 100 hours of meetings is collected with a variety of microphones, video cameras, electronic pens, presentation slide and whiteboard capture devices. For technical reasons the meetings in the corpus are formed by a group of four persons.

The first step for the analysis of this data is the processing of the raw audio-visual stream. This involves various challenging tasks. In the AMI project we address the audio-visual recognition problems by formulating seven questions:

1. What has been said during the meeting?
2. What events and keywords occur in the meeting?
3. Who and where are the persons in the meeting?
4. Who in the meeting is acting or speaking?
5. How do people act in the meeting?
6. What are the participants' emotions in the meeting?
7. Where or what is the focus of attention in meetings?

The audio-visual processing workpackage within the AMI project aims to develop algorithms that can automatically answer each of these questions from the raw audio-visual streams. The answers can then be used either directly during or after the meeting (e.g. in a meeting browser), or as an input for a higher level analysis (e.g. summarisation). In this article we describe the progress that has been made in the first two AMI project years towards the automatic recognition from audio-visual streams, and thus towards answering the questions. Each of the next chapters discusses algorithms, methods, and evaluation standards for one of the seven questions and summarises the experiences we made.

2 What Has Been Said During the Meeting?

Meetings are an audio visual experience by nature, information is presented for example in the form of presentation slides, drawings on boards, and of course by verbal communication. The latter forms the backbone of most meetings. The automatic transcription of speech in meetings is of crucial importance for meeting analysis, content analysis, summarisation, and analysis of dialogue structure. Widespread work on automatic speech recognition (ASR) in meetings started with yearly performance evaluations held by the U.S. National Institute of Standards and Technology (NIST) [27]. This work was initially facilitated by the collection of the ICSI meeting corpus [14]. Additional meeting resources were made available from NIST, Interactive System Labs (ISL) [4] and the Linguistic Data Consortium (LDC), and more recently, the AMI project[7].

The objectives for work in ASR in meetings are to develop state-of-the-art speech recognition technology for meeting transcription; to enable research into meeting relevant topics into ASR; to provide a common working base for researchers; and to enable downstream processing by providing automatically annotated and transcribed data. All of these objectives require a common and standardised evaluation scheme and unified testing procedures. For ASR in general evaluations by word error rate measurement according to NIST protocols is standard. A more critical issue is the task definition with respect to input media and objective system output.

To ensure that the technologies under development are state of the art we participated in international evaluations of ASR systems for meeting transcription [27]. World-leading research groups in ASR enter this competition which aims to provide a comparison between different approaches by provision of standardised common data sets, an evaluation schedule, and by organisation of a workshop to ensure information exchange. AMI has successfully competed in the NIST RT05s STT evaluations [27], yielding very competitive results on both conference meeting and lecture room transcription [10, 11]. AMI specific evaluations are performed on AMI data alone. As all microphone conditions are available for the complete corpus no special targeted sub-sets are defined. In the course of our next development cycle we will implement a larger development test set (planning of this set had an input on data collection) that will cover all aspects of the corpus in terms of meeting room, scenario and speaker coverage.

Table 1 shows WER results for the 2005 AMI meeting transcription system. The high deletion rate is a main contributor to the error rate. The associated results on rt05seval MDM are also shown in Table 1, again with relatively high deletion rates. Particularly poor performance on VT data has a considerable impact on performance (only two distant microphones).

In summary, in the last two years we have defined an evaluation framework that is generic, flexible, comparable, and that allows us to conduct research and development in a stable environment. We have built a system that is very competitive and performs exceptionally well on AMI data. We can focus our attention on extending our work to the full AMI corpus and the specific problems to be faced there. Further a research infrastructure is in place that allows all

Table 1. WER in % on rt05seval IHM, respectively MDM.

	TOT	Sub	Del	Ins	Fem	Male	AMI	ISL	ICSI	NIST	VT
rt05seval IHM	30.6	14.7	12.5	3.4	30.6	25.9	30.9	24.6	30.7	37.9	28.9
rt05seval MDM	42.0	25.5	13.0	3.5	42.0	42.0	35.1	37.1	38.4	41.5	51.1

partners to work on subtasks without the need to build large and labour-intensive systems or work on oversimplified configurations.

Our research results also give a better understanding of many interesting questions such as the distant microphone problem, the segmentation problem, the use of language, or the presence of many accents. Our investigations highlight where major improvements in performance can be obtained.

3 What Events and Keywords Occur in the Meeting?

Acoustic event and keyword spotting (KWS) are important techniques for fast access to information in meetings. Here we will concentrate on KWS: the goal is to find the keyword and its in speech data including its position and confidence. In AMI we compared three approaches to KWS. They are based on a comparison of two likelihoods: that of the keyword and the likelihood of a background model.

In the *acoustic approach*, phoneme-state posteriors are first estimated using a system based on neural networks with split temporal context [21]. The models of keywords are assembled from phoneme models and run against a background model (a simple phoneme loop). The difference of two log-likelihoods at the outputs of these models forms the score. It is advantageous to pre-generate the phoneme-state posteriors. The actual decoding is then very fast. We have further accelerated the decoding by pruning the phoneme-state posterior matrices by masking them using phoneme lattices discussed below. Then the decoding runs about 0.01 $\times$ RT on a Pentium 4 machine. *KWS in LVCSR lattices* greps the keywords in lattices generated by a large vocabulary continuous speech recognition system (LVCSR, Sect. 2). The confidence of each keyword is the difference of the log-likelihood of the path on which the keyword lays and the log-likelihood of the optimal path. The *KWS in phoneme lattices* is a hybrid approach. First, phoneme lattices are generated. This is in fact equivalent to narrowing the acoustic search space. The phonetic form of the keyword is then grepped in such lattices and the confidence of keywords is given by the acoustic likelihoods of individual phonemes, again normalised by the optimal path in the lattice.

A detailed description of the different systems, features, and a comparison of neural networks and GMMs in acoustic KWS can be found in [25]. Table 3 presents results of the three approaches on three test-sets. The sets are carefully defined on the ICSI meeting database [14]. While "Test 17" contains 17 common words, the sets "Test 1 and 10" concentrate on rare words occurring at most

Table 2. Comparison of Figure-of-Merit (FOM) measure (in %) of KWS approaches.

Test set	Acoustic	Word lattice	Phoneme lattice
Test 17	64.46	66.95	60.03
Test 10	72.49	66.37	64.1
Test 1	74.95	61.33	69.3

one, respectively ten times in the test set. The results confirmed our previous assumptions about the advantages and drawbacks of the different approaches:

LVCSR-KWS is fast (lattices can be efficiently indexed) and accurate, however only for common words. We see a clear degradation of performance for the sets "Test 1 and 10". We should take into account that less common words (such as technical terms and proper names) carry most of the information and are likely to be searched by the users. LVCSR-KWS has therefore to be completed by a method unconstrained by the recognition vocabulary. *Acoustic KWS* is relatively precise (the precision increases with the length of the keyword) and any word can be searched provided its phonetic form can be estimated. This approach is ideal for on-line KWS in remote meeting assistants, but even with the mentioned high speed of 0.01 $\times$ RT, it is not suitable for browsing *huge* archives, as it needs to process all the acoustic (or at least posterior probabilities) data. *Phoneme lattice KWS* is a reasonable compromise in terms of accuracy and speed. Currently, our work on indexing phoneme lattices using tri-phoneme sequences is advancing and preliminary results show a good accuracy/speed trade-off for rare words.

With the acoustic keyword spotter, an on-line demo system was implemented. This system uses a new closed-form based algorithm for speaker segmentation which takes into account time information of cross-correlation functions, values of its maxima, and energy differences as features to identify and segment speaker turns [16]. As for LVCSR spotting, it was completed by an indexation and search engine [8] and integrated into the AMI multimodal browser JFerret [30].

Future work includes improvement of the core algorithms and on KWS enhanced by semantic categories.

4 Who and Where are the Persons in the Meeting?

To browse meetings and relate different meetings to each other it is important to know, who was actually in the meeting. In this section we first address the problem of identifying persons and then track them through the meeting. Once identified, we aim to track the persons location through the meeting room. The location of each meeting participant at each time instance is rather uninteresting for a later comprehension of a meeting. It is very unlikely that a user will browse a meeting and ask for the "three dimensional coordinates of participant A at time instance 03:24:12". However, while usually not used directly, the correct coordinates of each person in the meeting are an essential input to various other

meeting analysis tasks, including the focus of attention (Sect. 8) and action recognition (Sect. 6). Furthermore these methods rely on very exact coordinates; wrong coordinates will lead to an error propagation, or in the worst case, to a termination of subsequent tasks. Thus determining the correct location of each meeting participant at each time in the meeting is a very crucial task.

An identification of meeting participants is possible from both the face and the voice. Here we'll concentrate on the face. During recent international competitions on face authentication [15], it has been shown that the discriminant approaches perform very well on manually localised faces. Unfortunately, these methods are not robust to automatic face localisation (imprecision in translation, scale and rotation) and their performance degrades. On the opposite, generative approaches emerged as the most robust methods using automatic face localisation. This is our main motivation for developing generative algorithms [6, 5]. For AMI we proposed to train different generative models, such GMMs, 1D-HMMs, and P2D-HMMs, using MAP training instead of the traditionally used ML criterion. Currently, we are evaluating the algorithms on a face verification task using the well-known BANCA benchmark database [3]. Our results show that generative models are providing better results than discriminant models. The best results are achieved by P2D-HMM. However, it should be noted that P2-HMMs are also much slower than GMMs. The algorithms have been developed as a machine vision package for a well-known open source machine learning library called Torch vision [26]. This package provides basic image processing and feature extraction algorithms but also several modules for face recognition.

For localisation and tracking of the meeting participants we developed, applied, and evaluated four different methods. To evaluate these methods we used the AMI AV16.7 corpus. It consists of 16 meeting room sequences of 1-4 minutes length with up to four participants, recorded from two camera perspectives. The sequences contain many challenging phenomena for tracking methods, like person occlusion, cameras blocked by passing people, partial views of backs of heads, and large variations in the head size. A common evaluation scheme, based on the procedure defined in [23] and a defined training and test corpus, allows to compare the advantages and the drawbacks of the different methods.

The *trans-dimensional MCMC* tracker is based on a hybrid Dynamic Bayesian Network that simultaneously infers the number of people in the scene and their body and head locations in a joint state-space formulation [22]. The method performs best when tracking frontal heads in the far field view. The *Active Shape* tracker is based on a double layered particle filter framework, which on the one hand allocates sets of particles on different skin coloured blobs and evaluates predicted head-shoulder contours on the image data. Especially in scenes with partial occluded heads the tracking algorithm shows its great performance. The *Kanade-Lucas-Tomasi (KLT)* tracking uses an image pyramid in combination with Newton-Raphson style minimisation to find a most likely position of features in a new image [13]. This method tracks heads correctly in more than 90% of the video sequences, however hands are often misinterpreted as heads. The *face detector* is based on a skin colour blob extraction followed by a movement

prediction. The face detector is based on the weak classifier compound of a Gabor wavelet and a decision tree [19]. The negative aspect of this face detector is the strong computation dependency on the Gabor wavelet feature evaluation and therefore it can not be used in real-time applications.

A comparative study of the four different head tracking methods, a detailed descriptions of the algorithms, and evaluation results can be found in [24].

5 Who in the Meeting is Acting or Speaking?

The objective of this work is to be able to segment, cluster and recognise the speakers in a meeting, based on their speech. Speaker information can be included in the meeting browser so that the user will have a better understanding of what is going on and will have a better context of the contents.

Within AMI we developed two approaches. The first uses the acoustic contents of the microphone signal to segment and cluster speakers. This extends earlier TNO work on speaker recognition (for telephone speech) and speaker segmentation/clustering (for broadcast news). The system has been evaluated in the NIST Evaluation on Meeting Data [27]. The evaluation set contained ten meetings in total, two meetings each from five different sources. One meeting source was AMI. We participated in both the Speech Activity Detection task and the Speaker diarisation task. The system obtained very competitive results in the NIST RT05s evaluation for speech activity detection (the lowest error rate reported) and our speaker diarisation system performed satisfactorily, given the technology we used.

The second system is a new closed-form localisation based algorithm which takes into account time information of cross-correlation functions, values of its maxima, and energy differences as features to identify and segment speaker turns [16]. In order to disambiguate timing differences between microphone channels caused by noise and reverberation, initial cross-correlation functions were time-smoothed and time-constrained. Finally we used majority voting based scoring approach to decide about the speaker turns.

The system was tested on challenging data recorded within the AMI project (ICSI, AMI-pilot, and BUT data) recorded at 16kHz. Achieved results show that the between-channel timing information brings sufficient information about speaker turns, especially in case of segmenting heavily cross-talked data. The achieved frame-level accuracy (for every 10ms) is around 90% for all three databases even thought the degree of the cross-talk (influencing the reliability of particular hypothesis) varies a lot between different meeting data.

The proposed system has been successfully applied to segment newly created real meeting-recordings for AMI. Obtained rough speaker-turns (with speech and silence segmentation based on classical MFCCs classified using neural network trained on ICSI training data set) are exploited by annotators to create word-level orthographic transcriptions of new AMI meeting data.

For demonstration purposes, we also developed a speaker segmentation system that is able to detect speaker turns in real time. The system has been

proposed together with acoustic based key-word spotter (Sect. 3). Furthermore, on-line pre-processing of visual input from the camera, scanning the whole scene using the hyperbolic mirror has been used.

6 How Do People Act in the Meeting?

We aim to extract visual features from videos and develop methods to use them for the automatic recognition of important actions and gestures in meetings. We focus on semantic actions and gestures that indeed happen in meetings and that can be of potential use to the user of a meeting-browser or as a cue for higher-level tasks in group analysis. We have defined a set of actions and gestures that are relevant for meetings, these include hand, body, and head gestures. Examples are Pointing, writing, standing up, or nodding. Special attention has been paid to negative signals, i.e. a negative response to a yes-no question usually characterised 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 the *gesture segmentation* two methods were applied: Bayes Information Criterion and an Activity Measure approach. As features we used Posio [18] (cf. Sect. 8) to extract 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. In addition we performed *classification of the segmented data*. Due to the temporal character of gestures we focused on different HMM methods.

The main conclusion regarding the automatic segmentation of gestures in real meetings is that it still a very challenging problem and the tested approaches do not give good segmentation performance for whole gestures, mainly due to the intrinsic structure of the gestures and the noise in the input features. An alternative approach is to develop segmentation algorithms for gesture parts and in preliminary evaluations this gave promising results.

Given this segmentation experience, the classification task was performed on manually segmented video streams. We found that a garbage model improves the recognition performance significantly. The HMM approaches gave a reasonable performance. Gestures like standing up (100% recognition rate) and the important speech supporting gestures (85%) reached results satisfactory for practical applications. However the results for the detection of negative signals were not significantly better than guessing. Detecting gestures such as shaking or nodding and negative signals is still a challenging problem that requires methods capable of detecting very subtle head movements.

In summary: important gestures and actions in meetings, such as negative signals are very hard to detect, as they can be very subtle. The standard algorithms used for artificial gestures – such as HMMs – can therefore not be applied directly to the meeting domain. Methods capable of detecting very small movements are required and have to be investigated in detail.

7 What are the Participants Emotions' in Meetings?

Recent studies [12] on emotions in meetings showed that people are – of course – not showing all kind of emotions in meetings, but only a rather small subset like bored, interested, serious, etc. On the other hand some emotions, like sadness, are very unlikely to appear. Furthermore peoples' expression of emotions in meetings is rather subtle compared to artificial emotion databases (see [17] for a recent survey). The combination of these two fact makes the detection of emotions in meetings rather difficult and calls directly for special methods. Similar to our AMI experience with gestures and actions (Sect. 6) standard methods for emotion detection from acted databases can not be directly applied to meetings.

AMI therefore aims to develop special algorithms to estimate the meeting participants' emotion from the information of head- and body pose, gestures and facial expressions. Therefore, the development and enhancement of the corresponding algorithms is crucial for emotion recognition by visual input. A description of activities can be found in Sect. 6. Independently, works are going on to analyse facial expressions. Very recent investigations are based on an application of the AdaBoost [9] algorithm and its variants applied on two-dimensional Haar- and Gabor-Wavelet coefficients, for localisation of frontal faces and eyes [28], as well as for classification of facial expressions [17]. Furthermore, an approach based on Active Appearance Models is implemented and investigated in its application to head pose estimation and facial expression analysis.

Evaluation of these – especially to the meeting domain adapted algorithms – is currently ongoing, showing very promising results. Even though this method shows high requirements to the computational performance of the applied hardware, the expected results argue for this approach.

8 Where or What Is The Focus of Attention in Meetings?

There are two questions to answer when trying to understand what is going on during the meeting. However, in view of the difficulty to determine both the group focus of attention (FOA) and the general FOA of individual people (a person might have multiple FOA – listening to a speaker while taking notes –, ground truthing a mental state is difficult), we restricted our investigations to the visual FOA of people defined as the spatial locus defined by the person's gaze, which is indeed one of the primary cue for identify the attentional state of someone [20]. With this definition, research was conducted into two directions.

In the first direction, the objective is to identify the role played by the FOA in the dynamics of meetings. Answering such questions will be useful to understand the relationship between the FOA and other cues (such as speaker turns, cf. Sect. 5) as well as to more precisely identify the interactions between participants (e.g. by contributing to the recognition of the higher level dialog acts), which in turn could translate to better FOA recognition algorithms. The second direction is concerned with the recognition of the FOA. More precisely, given recorded meeting data streams, can we identify at each instant the FOA of the meeting participants? Both directions were investigated and are summarised in four tasks.

Perception of head orientation in a Virtual Environment: This task consists of assessing how accurately people perceive gaze directions. In a virtual environment an avatar was positioned at one side of the table. At the other side a number of balls where placed at eye height for the avatar. Persons where then asked to predict at which ball the avatar was looking at. As a first result we found that there is no significant difference for the location of the avatar. Furthermore no learning effect among the participants has been found. Decreasing the angle between the balls increases the judgement error. With an azimuth angle between two persons at one side of the table of 30 degree, as seen from a person at the other side, an discrimination is possible with an accuracy of 97.57%. This shows that head orientation can be used as a cue for the FOA.

Identifying speaker amongst meeting participants: In this task AMI investigates, whether observers use knowledge about differences in head orientation behaviour between speakers and listeners by asking them to identify the speaker in a four-person setting. In a thorough study on the role of FOA in meeting conversations, we showed through the use of a Virtual Environment display that people are indeed using the gaze and head pose of participants to assess who is speaking. This results demonstrate that humans apply knowledge about systematic differences in head orientation behaviour between speakers and listeners. This shows how important the FOA in meetings is.

Head pose and head tracking: (cf. Sect. 4) One first step towards determining a person's FOA consists of estimating its gaze direction. Then from the geometry of the room and the location of the participants, the FOA can normally be estimated. However, as estimating gaze is difficult (and requires very close-up views of people to assess the position of the pupil in the eye globe), AMI has developed, as an approximation, algorithms for tracking the head and estimate its pose. We formulate the coupled problems of head tracking and head pose estimation in a Bayesian filtering framework, which is then solved through sampling techniques. Details are given in [2, 1]. Results were evaluated on 8 minutes of meeting recordings involving a total of 8 people, and the ground truth was obtain from flock-of-birds (FOB) magnetic sensors. The results are quite 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.

Recognition of the FOA: In this task, the emphasis is on the recognition of a finite set $\mathcal{F}$ of specific FOA loci. Unlike other works, the set of labels in our setting was not restricted to the other participants, but included also looking at the table (e.g. when writing), at a slide screen, and an unfocused label (when looking at any direction different than those of the other labels). One approach to the FOA recognition problem that we have followed consists of mapping the head pose orientations of individual people to FOA labels. This was done by modelling each FOA with a Gaussian and the unfocus class with a uniform distribution. Evaluation was conducted on 8 meetings of 8 minutes on average. Each meeting involved 4 people, and the FOA of two of them was annotated.

First, we conducted experiments by using the head-pose pointing vectors obtained from the ground truth FOB readings. We obtained a frame-based classification rate of 68% and 47%, depending on the person's position in the smart meeting room. These numbers are lower than those reported in other works, and are mainly due to the use of a more complex setting, more labels, and demonstrate the impact of FOA spatial configurations on the recognition, and the necessity of exploiting other features/modalities (e.g speaking status) in addition to the head pose to disambiguate FOA recognition. Furthermore we found that using the estimated head-pose instead of the ground truth were degrading the results not so 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 adaption approaches like Maximum A Posteriori techniques for the FOA recognition. These adaptation techniques, along with the use of multimodal observation, will be the topic of current research.

9 Conclusion

In this article we described how audio-visual processing in meeting scenarios can be addressed with seven questions. We showed, how the project AMI developed and applied machine learning techniques to answer each of the questions automatically. By addressing the different audio-visual tasks with simple questions we were able to streamline and coordinate the development process and enable an easy sharing of data and recogniser outputs among the involved partners. This led to common evaluation schemes on commonly defined AMI data sets for each of the tasks and allows us to compare different approaches in a simplified way. Finally it is worth to mention, that this has been achieved by more than 50 persons from eight institutes in seven countries in the EU and the US.

References

[1] S.O. Ba and J.M. Odobez. Evaluation of head pose tracking algorithm in indoor environments. In *Proceedings IEEE ICME*, 2005.

[2] S.O. Ba and J.M. Odobez. A rao-blackwellized mixed state particle filter for head pose tracking. In *Proceedings of the ACM-ICMI Workshop on MMMP*, 2005.

[3] BANCA. Benchmark database. `http://www.ee.surrey.ac.uk/banca`.

[4] S. Burger, V. MacLaren, and H. Yu. The ISL meeting corpus: The impact of meeting type on speech style. In *Proceedings ICSLP*, 2002.

[5] F. Cardinaux, C. Sanderson, and S. Bengio. Face verification using adapted generative models. In *Int. Conf. on Automatic Face and Gesture Recognition*, 2004.

[6] F. Cardinaux, C. Sanderson, and S. Marcel. Comparison of MLP and GMM classifiers for face verification on XM2VTS. In *Proc. IEEE AVBPA*, 2003.

[7] J. Carletta et al. The AMI meetings corpus. In *Proc. Symposium on Annotating and measuring Meeting Behavior*, 2005.

[8] M. Fapso, P. Schwarz, I. Szoke, P. Smrz, M. Schwarz, J. Cernocky, M. Karafiat, and L. Burget. Search engine for information retrieval from speech records. In *Proceedings Computer Treatment of Slavic and East European Languages*, 2005.

[9] Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, 1996.
[10] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals. The 2005 AMI system for the transcription of speech in meetings. In *Proc. of the NIST RT05s workshop*, 2005.
[11] T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, and S. Renals. Transcription of conference room meetings: an investigation. In *Proceedings Interspeech*, 2005.
[12] D. Heylen, A. Nijholt, and D. Reidsma. Determining what people feel and think when interacting with humans and machines: Notes on corpus collection and annotation. In J. Kreiner and C. Putcha, editors, *Proceedings 1st California Conference on Recent Advances in Engineering Mechanics*, 2006.
[13] M. Hradis and R. Juranek. Real-time tracking of participants in meeting video. In *Proceedings CESCG*, 2006.
[14] 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 *Proceedings IEEE ICASSP*, 2003.
[15] K. Messer, J. Kittler, M. Sadeghi, M. Hamouz, A. Kostyn, S. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, N. Poh, Y. Rodriguez, and J. Czyz et al. Face authentication test on the BANCA database. In *Proceedings ICPR*, 2004.
[16] P. Motlicek, L. Burget, and J. Cernocky. Non-parametric speaker turn segmentation of meeting data. In *Proceedings Eurospeech*, 2005.
[17] M. Pantic and L.J.M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE TPAMI*, 22(12):1424–1445, 2000.
[18] R. Poppe, D. Heylen, A. Nijholt, and M. Poel. Towards real-time body pose estimation for presenters in meeting environments. In *Proceedings WSCG*, 2005.
[19] I. Potucek, S. Sumec, and M. Spanel. Participant activity detection by hands and face movement tracking in the meeting room. In *Proceedings CGI*, 2004.
[20] R. Rienks, R. Poppe, and D. Heylen. Differences in head orientation for speakers and listeners: Experiments in a virtual environment. *Int. Journ. HCS*, to appear.
[21] P. Schwarz, P. Matějka, and J. Černocký. Hierarchical structures of neural networks for phoneme recognition. In *Accepted to IEEE ICASSP*, 2006.
[22] K. Smith, S. Ba, J. Odobez, and D. Gatica-Perez. Evaluating multi-object tracking. In *Workshop on Empirical Evaluation Methods in Computer Vision*, 2005.
[23] K. Smith, S. Ba, J.M. Odobez, and D. Gatica-Perez. Multi-person wander-visual-focus-of-attention tracking. Technical Report RR-05-80, IDIAP, 2005.
[24] K. Smith, S. Schreiber, V. Beran, I. Potúcek, and D. Gatica-Perez. A comparitive study of head tracking methods. In *MLMI*, 2006.
[25] I. Szöke, P. Schwarz, P. Matějka, L. Burget, M. Karafiát, M. Fapšo, and J. Černocký. Comparison of keyword spotting approaches for informal continuous speech. In *Proceedings Eurospeech*, 2005.
[26] Torch. `http://www.idiap.ch/~marcel/en/torch3/introduction.php`.
[27] NIST US. Spring 2004 (RT04S) and Spring 2005 (RT05S) Rich Transcription Meeting Recognition Evaluation Plan. Available at `http://www.nist.gov/`.
[28] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 2002.
[29] A. Waibel, H. Steusloff, R. Stiefelhagen, and the CHIL Project Consortium. CHIL: Computers in the human interaction loop. In *Proceedings of the NIST ICASSP Meeting Recognition Workshop*, 2004.
[30] P. Wellner, M. Flynn, and M. Guillemot. Browsing recorded meetings with Ferret. In *Proceedings MLMI*. Springer Verlag, 2004.