Recognition and Understanding of Meetings The AMI and AMIDA Projects

Steve Renals [1] Thomas Hain [2]
Hervé Bourlard [3]

IDIAP–RR 07-46

October 4, 2007

to appear in
Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU'07

[1] Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH3 5EU, UK, s.renals@ed.ac.uk

[2] Dept. of Computer Science, University of Sheffield, Sheffield S1 4DP, UK, t.hain@dcs.shef.ac.uk

[3] IDIAP Research Institute, CP 592, 1920 Martigny, Switzerland and Ecole Polytechnique Fdrale de Lausanne (EPFL), Switzerland, herve.bourlard@idiap.ch

IDIAP Research Report 07-46

Recognition and Understanding of Meetings The AMI and AMIDA Projects

Steve Renals Thomas Hain Hervé Bourlard

October 4, 2007

to appear in

Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU'07

Abstract. The AMI and AMIDA projects are concerned with the recognition and interpretation of multiparty meetings. Within these projects we have: developed an infrastructure for recording meetings using multiple microphones and cameras; released a 100 hour annotated corpus of meetings; developed techniques for the recognition and interpretation of meetings based primarily on speech recognition and computer vision; and developed an evaluation framework at both component and system levels. In this paper we present an overview of these projects, with an emphasis on speech recognition and content extraction.

Contents

1 Introduction 3

2 The AMI Corpus 4

3 Audio-video Processing 4

4 Meeting Speech Recognition 5
4.1 Meeting domains 5
4.2 Audio Pre-processing 7
4.2.1 Individual Head Microphones 7
4.2.2 Multiple Distant Microphones 8
4.3 Acoustic modelling 8
4.4 Language modelling 9
4.5 Performance 9

5 Content Extraction 10
5.1 Dialogue act recognition 11
5.2 Topic segmentation 12
5.3 Summarization 12

6 Evaluation 12

7 Conclusions and future work 13

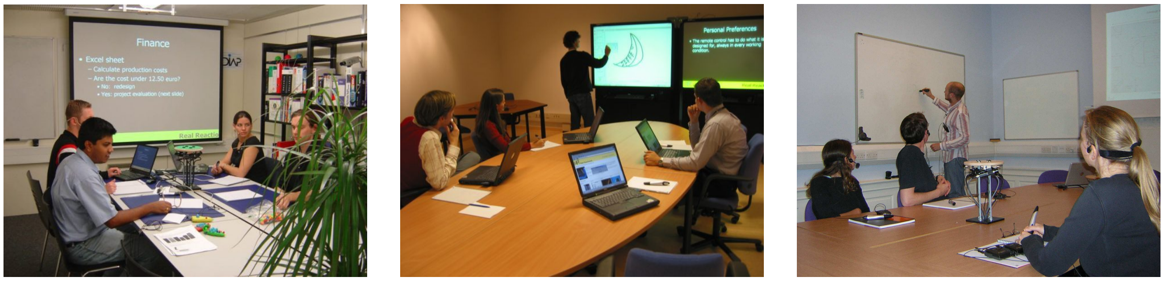

Figure 1: The three AMI instrumented meeting rooms at IDIAP (left), TNO (centre) and the University of Edinburgh (right).

1 Introduction

In recent times there has been growing research interest in the recognition and understanding of interactions between people in settings such as meetings, lectures, seminars and teleconferences. The modelling and interpretation of human-human communication scenes is a challenging scientific endeavour, requiring a broad range of research advances in areas including signal processing, speech recognition, multimodal scene analysis, discourse analysis, and multimodal retrieval. The analysis and interpretation of multiparty meetings is of scientific interest since it provides a circumscribed arena for the investigation of communication scenes, as well as underpinning a number of potentially significant applications.

Meetings play a crucial role in the generation of ideas, documents, relationships, and actions within an organization. The wealth of information exchanged in meetings is often lost, at least in part, because human note taking of meeting minutes is subjective and incomplete, capturing only a fraction of the information. Multimodal recording of meetings is an attractive alternative, but such recordings will only become really useful once it is possible to recognize, structure, index and summarize them automatically.

Since the mid-1990s a number of resarchers have investigated the automatic recording, recognition and interpretation of meetings [1, 2, 3, 4, 5, 6, 7]. From 2004, the AMI consortium[1], has investigated the development of technologies to enhance human collaboration in the domain of meetings. AMI is concerned with the development of algorithms, models, and prototype systems that support interaction in meetings and access to meeting-related information. Our initial research was concerned primarily with the analysis of face-to-face meetings recorded in an instrumented meeting room equipped with multiple microphones and cameras, and capturing other interaction modalities including the handwriting and data projected slides. More recently we have extended the focus of our work to support meetings where some of the participants may be remote, and to provide services to operate on meetings both in realtime and on an archive.

Much of the research that we have carried out has built on a corpus of 100 hours of multimodal meeting recordings annotated at a number of different levels, outlined in Section 2. Some of the core work of the AMI consortium has been the development of recognizers for audio and video modalities, including gesture and action recognition and audio-visual tracking. These are briefly outlined in Section 3, which is followed by a more detailed discussion of the AMI system for automatic speech transcription of meetings, from both close-talking and distant microphones (Section 4). The output of the multimodal recognizers, in particular the automatic speech transcription, forms the basis of our work in content extraction, including topic segmentation, summarization and dialogue act recognition, discussed in Section 5. A key aspect of our work has been a focus on evaluation, both at the component and system levels (see Section 6), the latter being closely tied to the design of the AMI corpus.

[1] This work has been primarily carried out in the context of two EU Integrated Projects AMI, and its follow-on project AMIDA (`http://www.amiproject.org/`).

2 The AMI Corpus

Much of our research is built on the use of instrumented meeting rooms to collect recordings of multiparty meetings. Three standardized meeting rooms were designed and constructed at AMI partners IDIAP, TNO and the University of Edinburgh (Figure 1). These rooms, which were designed for the collection of four person meetings, all contained a set of standardized recording equipment:

- six cameras — four providing close-up views of the participants, two providing a view of the whole room;
- twelve microphones — a headset microphone per participant and an 8-element circular microphone array;
- data projector capture (VGA);
- whiteboard capture;
- digital pen capture.

There were also additional recording devices in each of the rooms, including an additional microphone, a binaural manikin and additional cameras.

These instrumented meeting rooms were used to record the AMI Meeting Corpus [8], which consists of 100 hours of meeting recordings, with the different recording streams synchronized to a common timeline. The corpus includes manually produced orthographic transcriptions of the speech used during the meetings, aligned at the word level. In addition to these transcriptions, the corpus includes manual annotations that describe the behaviour of meeting participants at a number of levels. These include dialogue acts, topic segmentation, extractive and abstractive summaries, named entities, limited forms of head and hand gestures, gaze direction, movement around the room, and where heads are located on the video frames. Not all 100 hours of meetings have been marked with all kinds of annotations. The linguistically motivated annotations have been applied most widely, covering at least 70% of the corpus in all cases. The annotations were carried out using NXT (the NITE XML Toolkit) [9], an open source XML-based infrastructure for the annotation and management of multimodal recordings[2].

The corpus consists of two types of meetings: a design scenario, and naturally occurring meetings in a range of domains. About 70% of the corpus was elicited using the scenario in which the participants play different roles in a design team, taking a design project from kick-off to completion over the course of a day. The scenario meetings consist of a series of four meetings, attended by four participants, who had tasks to accomplish between meetings. The participant roles were driven in real-time by emails and web information. There are several advantages to recording scenario meetings. First, it enabled us to control the domain, making it easier to understand the content of the meetings, and to enable the construction of deeper approaches to content extraction. Second, the construction of a meeting scenario enabled outcome measures to be defined, including preferred design outcomes (not achieved by any of the teams!). Third, the fact that participants were not part of a real organization made it much easier to control their knowledge and motivation. Fourth, scenario meetings are replicable, and thus enable system-level evaluations, such as the task-based evaluation discussed in Section 6.

The corpus is publicly available on the web at `http://corpus.amiproject.org`, and is released under a licence that is based on the terms of the Creative Commons Attribution NonCommercial ShareAlike 2.5 Licence.

3 Audio-video Processing

The construction of audio-video recognizers is at the heart of the automatic processing of multimodal meeting records. In the AMI project we have developed a number of recognizers for the multimodal meeting recordings, including speech transcription (discussed below), speaker diarization [10], audio-video localization and

[2] `http://sourceforge.net/projects/nite/`

tracking [11], and visual focus of attention [12]. The outputs of these recognizers may be used directly, e.g. in a meeting browser, or as input for some higher level analysis (Section 5).

4 Meeting Speech Recognition

Raw transcription of the spoken conversations in meetings is essential for most types of higher level analysis such as content extraction (Section 5). Work on meeting transcription in the past was dominated by the fact that the amount of in-domain data was very small. As the type of speech is conversational in most cases, the cost of manual transcription for the purpose of model training is usually prohibitive. The amount of speech resources for meetings is still relatively small and most systems make use of adaptation of models from other domains. In [13], a recognition system for conversational telephone speech (CTS) formed the starting point, others have reported that bootstrapping from Broadcast News (BN) systems works well.

Design of a transcription system requires appropriate definition of the domain. This is especially difficult as any gathering of people discussing arbitrary topics could be interpreted as a “meeting”. Hence suitable constraints need to be defined. The U.S. National Institute of Standards and Technology (NIST) has in the past made the distinction between two types of meetings: conference room meetings where people gather around a table to discuss multiple topics, following a certain agenda; and lecture room meetings where a single speaker presents to an audience, but may also engage in discussion with the audience. AMI meetings belong mostly to the first category, however occasionally a participant stands up to give a presentation, hence fitting into the second type. These distinctions have significant implication on acoustics and language use.

Another issue, independent of such classification, is the recording source variability. Most meeting corpora have audio recorded with individual head microphones (IHM). However, for convenience, ideally only microphones on the table, in microphone array configuration or free standing, should suffice for this task (multiple distant microphones, MDM). Naturally though, because of occlusion, noise and reverberation, for MDM data a substantial performance degradation in recognition can be observed.

The AMI transcription system makes use of a standard ASR framework employing hidden Markov model (HMM) based acoustic modeling and N-gram based language models (LMs). In the following we briefly address issues of the domain, followed by a brief description of the essential components of a meeting transcription system and the performance in recent NIST evaluations. For a more elaborate description of the systems, the interested reader is referred to [14].

4.1 Meeting domains

Even within the set of conference room meetings the recordings can vary considerably. Apart from the AMI corpus several other meeting corpora are available: the ICSI Meeting corpus [15], two phases of the NIST corpus [16], and the ISL recordings [17]. In addition, recordings from Virginia Tech (VT) and the European project CHIL have been used in NIST evaluations. In Table 1 below the raw average segment statistics for these corpora are compared. A segment here is defined as speech not interrupted with silence of at least 100ms length. As can be observed, segment lengths vary greatly with the AMI corpus recordings having much longer sentences on average, hinting at speech at a more controlled pace. The very short segments on the CHIL data are surprising, given that these recordings belongs to the lecture room meetings. The speaking rate however is very similar for all corpora, varying between 3.1 and 3.6 words per second.

Apart from these raw statistics, the acoustics and language can differ. Acoustic variation is mostly given by the recording setup (see the following section), but the language situation is less clear. The aforementioned meeting corpora differ not only in recording configuration but also in topics and the style of discussion. In the first instance one can look at vocabulary differences for the different corpora. Table 2 shows out of vocabulary (OOV) rates using vocabulary derived from each meeting corpus. The OOV rates do not correlate perfectly with vocabulary sizes in these corpora and overall the mismatch of ISL vocabulary to the other corpora is greatest.

Table 3 shows the same analysis as before, however in this case the word lists are padded with the most frequent words from BN texts to yield 50k words. It is evident that overall the effect of vocabulary mismatch is greatly reduced for all cases. This suggests that only a very small amount of meeting specific vocabulary

Meeting resource	Avg Dur (sec)	Avg. Words/Seg
ICSI	2.11	7.30
NIST	2.26	7.17
ISL	2.36	8.77
AMI	3.29	10.09
VT	2.49	8.27
CHIL	1.80	5.63

Table 1: Segment statistics for meeting corpora.

	Vocabulary Source			
Corpus	ICSI	NIST	ISL	AMI
ICSI	0.00	4.95	7.11	6.83
NIST	4.50	0.00	6.50	6.88
ISL	5.12	5.92	0.00	6.68
AMI	4.47	4.39	5.41	0.00
ALL	1.60	4.35	6.15	5.98

Table 2: %OOV rates of meeting resource specific vocabularies. Columns denote the word list source, rows the test domain.

	Vocabulary Source			
Domain	ICSI	NIST	ISL	AMI
ICSI	0.01	0.47	0.58	0.57
NIST	0.43	0.09	0.59	0.66
ISL	0.41	0.37	0.03	0.57
AMI	0.53	0.53	0.58	0.30
ALL	0.16	0.42	0.53	0.55

Table 3: %OOV rates of padded vocabularies. Columns denote the word list source, rows the test domain.

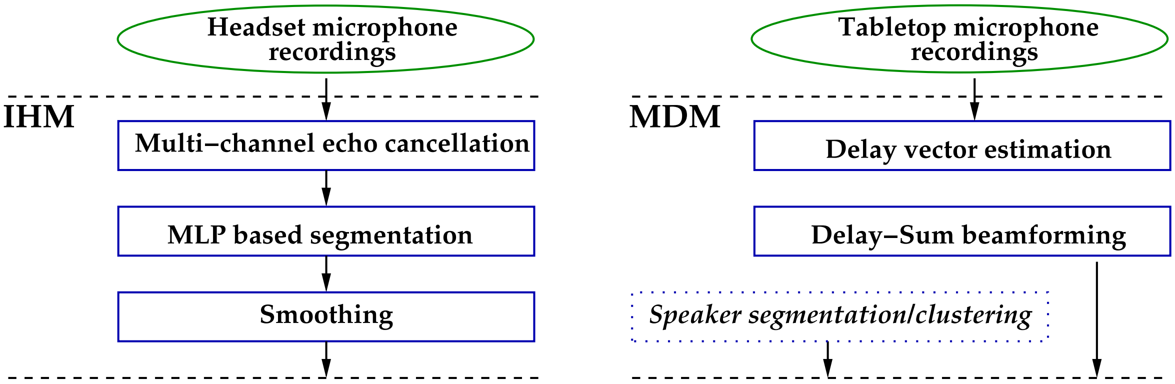
Headset microphone recordings
IHM
Multi−channel echo cancellation
MLP based segmentation
Smoothing
Tabletop microphone recordings
MDM
Delay vector estimation
Delay−Sum beamforming
Speaker segmentation/clustering

Figure 2: Front-end processing stages for IHM and MDM.

is necessary. In [18], an equivalent set of experiments was conducted using meeting room specific language models with padded vocabularies. Here the picture is less clear even though the corpus specific language models always give the lowest perplexity. However, the use of such language models in decoding does not give any gain over combined models.

4.2 Audio Pre-processing

The audio pre-processing stages have to address several issues: The segmentation of the audio and implicit discarding of silence or noise; the speaker labeling for later acoustic adaptation; the normalisation of input channels; and the suppression of noise. Audio can come from different microphone sources associated with a person (head mounted, lapel) or microphones in the relative vicinity (table, wall mounted). This separation is equivalent to the NIST distinction between IHM and MDM channels which implicitly groups recordings with substantially different acoustic properties.

IHM sources obviously carry implicit speaker information, hence the main task is segmentation. Whereas head mounted sources mostly acquire sounds from one speaker, lapel microphone recordings suffer from high levels of crosstalk and occlusion. For MDM the location of the microphones in the room relative to the speakers as well as relative to each other is important. Multiple microphones in array configuration allow substantially better performance than those in arbitrary location or a single microphone. Room reverberation can increase degradation compared with IHM results.

Most meeting transcription systems try to set up the front-ends such that later processing is identical, regardless whether audio comes form IHM or MDM sources. The AMI system is no exception to that (Figure 2); in particular the audio is enhanced and a single audio stream is presented to the recognition stages where the signal is converted into a single feature stream.

4.2.1 Individual Head Microphones

Initially cross-talk suppression is performed using an adaptive LMS echo canceller followed by computation of 12 MF-PLP features. Additional features for the detection of cross-talk are extracted prior to cross talk suppression. These features are cross-channel normalised energy, signal kurtosis, mean cross-correlation and maximum normalised cross-correlation. The cross-channel normalised energy is calculated as the energy for the present channel divided by the sum of energies across all channels. The feature vectors are used to train a multi-layer perceptron (MLP) classifier with a 101 frame input layer, a 50 unit hidden layer and an output layer of two classes. The models are trained on 90 hours of data from all conference room meetings resources available. For example, on the NIST RT'05 evaluation set (*rt05seval)* the automatic segmentation gave equal performance to manual segmentation in terms of WER. More details can be found in [19, 20].

System	Training criterion	PLP	LCRC+PLP
Baseline	ML	28.7	25.2
SAT	ML	27.6	23.9
SAT	MPE	24.5	21.7

Table 4: %WER results on *rt05seval* IHM (manual segmentation) with and without LCRC features.

4.2.2 Multiple Distant Microphones

For MDM a commonly used enhancement based approach was adopted where multiple channels were converted into a single one consisting of the dominant speaker only. Note that this approach cannot cope with speech from multiple speakers at the same time. Figure 2 shows the processing steps in diagrammatic form. Processing of MDM data takes account of the varying number of microphone channels and potentially unknown location of microphones in relation to each other (to allow for comparison beyond the AMI corpus). The processing operates in several stages: First gain calibration is performed by normalising the maximum amplitude level of each of the input files. Then the background noise spectrum is estimated using the lowest energy frames in the recording and this is used to Wiener-filter the data to remove stationary noise. In the next step delay vectors between channels are calculated on a per frame basis using generalised cross-correlation. Delays are computed in relation to a reference channel which also yields a relative scale factor. Delays and scale factors are then used in the final stage implementing superdirective beam-forming. More details can be found in [21].

While this approach is robust to a variety of configurations, for a small number of sparsely located microphones the estimates are unreliable. In this case simply selecting the channel with the highest energy for every time frame was found to yield substantially lower word error rates.

4.3 Acoustic modelling

As mentioned above, the amount of data available for acoustic model training from meetings is still relatively small compared to other domains such as CTS. Hence in the most recent AMI systems models trained on 2000 hours of CTS data are adapted with 170 hours of meeting data.

12 MF-PLP[3] features are extracted at a rate of 100 Hz and together with the zeroth cepstral coefficient form the basic feature vector. First and second derivatives are added. More recently the standard systems augment this feature vector with 25 phoneme posterior derived components. These so-called *left context – right context* (LCRC) features [22] are derived from multiple stages of MLPs that try to estimate phoneme state posterior probabilities. The input to these is not only the feature vector at the current time, but 25 surrounding frames as well.

All acoustic models employ cross-word state-clustered triphone models. It was found that, similar to CTS, 10–15% relative WER gain can be obtained using maximum likelihood based vocal tract length normalisation (VTLN) [23]. Secondly, heteroscedastic linear discriminant analysis (HLDA) gives consistent performance improvements [23]. Further gains can be obtained by discriminative training based on the minimum phone error (MPE) criterion [24], also jointly with constrained maximum likelihood regression (MLLR) based speaker adaptive training (SAT). The left column of Table 4 shows WER results for models trained on 100 hours of meeting data and the *rt05seval* test set. In both cases substantial improvements are found.

Adapting to the meeting domain from CTS raises the issue that CTS operates on audio with reduced bandwidth. In [23], it was shown that better performance can be obtained using the full bandwidth available. As a consequence, an MLLR based transform from narrow-band to wide-band data was derived and used in MAP adaptation of CTS models to meeting data. However, such a scheme is not viable with both HLDA and discriminative training. A solution to this problem is the projection of the meeting data into a narrowband space where both HLDA statistics can be gathered and discriminative training can be performed without regeneration of training lattices.

[3] This is the standard implementation of perceptual linear coefficients in the Hidden Markov Model Toolkit (HTK)

Initial models	Adaptation	WER
CTS SAT MPE		30.4
CTS SAT MPE	ML-MAP	26.0
CTS SAT MPE + ML-MAP	MPE-MAP	23.9

Table 5: %WER results on *rt05seval* IHM with adaptation from CTS

LM component	size	weights (trigram)
AMI data (prelim.)	206K	0.038
Fisher	21M	0.237
Hub4 LM96	151M	0.044
ICSI meeting corpus	0.9M	0.080
ISL meeting corpus	119K	0.091
NIST meeting corpus	157K	0.065
Switchboard/Callhome	3.4M	0.070
webdata (meetings)	128M	0.163
webdata (fisher)	128M	0.103
webdata (AMI)	138M	0.108

Table 6: Language model data set sizes and weights in interpolation.

Initial full covariance statistics are estimated on the CTS training set. A single constrained MLLR transform is estimated to map the 52D wideband (WB) meeting data to a 52D narrowband (NB) CTS space. The meeting data is mapped with this transform and full covariance statistics are obtained using models based on CTS state clustering. The two sets of statistics are then combined using MAP and the combined set of statistics is used to obtain a joint HLDA transform (JT). Now combined models in JT space can be trained using both CTS and mapped meeting data. These are then used to retrain CTS models in JT space, followed by SAT and MPE training. Equivalent to adaptation of maximum likelihood trained models with MAP, the JT/SAT/MPE models are adapted to meeting data using MPE-MAP [25]. The inclusion of SAT requires the presence of speaker transforms on meeting data. These are obtained from SAT training of MAP adapted CTS models in JT space. Table 5 shows results in JT space. A comparison of the final WER results with that in Table 4 shows a 0.6% absolute improvement. A more detailed analysis of this procedure can be found in [26].

4.4 Language modelling

The main issue in language modelling is the acquisition of suitable data. In the AMI systems standard *n*–gram models up to 4$^{\text{th}}$ order are interpolated from models trained on many sources. Table 6 lists the most relevant text resources used for training. One can see that the amount of meeting data available is small and collection of data from the internet is known to yield significant improvements in perplexity and WER. Especially for meetings where the topic still may be new, it is important that such collection can be carried out efficiently and on potentially small amounts of data. Note that in contrast to other web-data, the AMI web-data was collected using techniques that target text that is different to the already existing background material [27]. From the interpolation weights it is clear that conversational data is most important. The perplexity of the interpolated was 84.3 for the interpolated trigram and 81.2 for the 4-gram model on the NIST RT'06 evaluation test set.

4.5 Performance

The complete AMI system for the transcription of meeting as used in the NIST RT'07 evaluations operates in a total of 10 passes. The initial pass only serves to obtain a rough transcript to provide input to adaptation with VTLN, SAT, and MLLR. The following passes then generate bigram word lattices which are expended using 4-gram language models and rescored using models that are differently trained, for example on meeting data only, or adapted models, or different configurations in the feature extraction. Since a detailed description of

Description	Tot	CMU	AMI	NIST	VT
Initial decode	37.4	47.7	29.3	33.8	38.4
Adapted	28.2	37.9	21.9	24.6	27.9
Best single output	25.4	34.5	20.4	21.1	25.3
Combined	24.9	33.9	19.8	20.9	24.7

Table 7: %WER results on IHM data of the AMI 2007 system on the NIST RT'07 evaluation set.

Description	Total	Sub	Del	Ins
Initial	44.2	25.6	14.9	3.8
Adapted	38.9	18.5	16.8	3.5
Final	33.7	20.1	10.7	2.9
Final - Man, Segments	30.2	18.7	9.4	2.0

Table 8: %WER results on MDM data of the AMI 2007 system on the NIST RT'07 evaluation set.

the system would go beyond of the scope of this paper, the interested reader is referred to [28]. Table 7 shows details for various stages in the system, from the initial decoding with unadapted models to the output of the best branch in the system. The outputs of several branches then can be combined, yielding the lowest word error rate. Data in this test set are taken from four different corpora. The substantial difference in performance between these data sets mostly originates from a different quality of microphones, even though heavily accented speech plays a role.

Table 8 shows results on the same data, obtained by using MDM input and a less complex system structure. One can observe that the difference in the initial pass between IHM and MDM recordings is 7% WER absolute which remains up to the final pass. Whereas the difference between the manual and automatic segmentation of data on IHM was found to give only 1.3%, it can be observed that for MDM the difference is 2.5%.

The results above were obtained with a system that was specifically trained on multiple meeting sources. In contrast experiments were conducted to produce automatic transcripts for the complete AMI corpus. The corpus was split into five parts of approximately equal size to allow training on four parts and testing on the fifth part. Acoustic models, dictionaries and trigram LMs (interpolated with background LMs) hence were derived from approximately 80 hours and then used to transcribe 20 hours of data each. So far only maximum likelihood training and standard MF-PLP features were used in the experiments, however with automatic and manual segmentation. Table 9 shows results for the complete corpus. Results for manual and automatic segmentation differ by 1.9% WER absolute initially and the difference slightly increases with adaptation even though the absolute WER level drops by approximately 6% absolute.

5 Content Extraction

The extraction of content from multimodal meeting recordings is largely based on the results of the audio-video processing described above. To achieve accurate content extraction from meeting recordings, our emphasis has been on models and algorithms that combine modalities. Automatically extracted content enables meetings to be indexed and structured at a semantically richer level than is possible using the raw output of the audio-video recognizers. Much existing work in this area is concerned with the extraction of content from written language; a major focus of AMI has been the extension of textual approaches to multimodal settings, involving the use of prosodic, video and contextual features.

Our work in this area has included the development of automatic approaches to the segmentation and classification of phenomena such as dialogue acts [29], topics [30], and dominance and influence [31], as well as abstractive and extractive summarization [32] and content-based automatic camera selection [33]. Using the AMI corpus for all tasks, we have been able to agree on evaluation measures and procedures that allow us to compare different approaches and techniques, both internally and externally.

Training	Segment.	Adapt.	Total
	Man		43.2
VTLN, HLDA	Man	VTLN	39.4
VTLN, HLDA	Man	VTLN, MLLR	36.8
	Auto		45.1
VTLN, HLDA	Auto	VTLN	41.2
VTLN, HLDA	Auto	VTLN, MLLR	38.9

Table 9: %WER results on the complete AMI corpus using maximum likelihood trained acoustic models with automatic (Auto) or manual (Man) segmentation, and adaptation with VTLN and/or MLLR.

Here we focus on our advances in three areas: dialogue act recognition, topic segmentation, and summarization.

5.1 Dialogue act recognition

Dialogue acts are labels for utterances which roughly categorize the speaker's intention. They are useful for various purposes in a dialogue or meeting processing situation, such as part of a browser which highlights all points where a suggestion or offer was recognized. However, dialogue acts also serve as elementary units, upon which further structuring or discourse processing may be based. For example, the summarization components that we have developed are based on the dialogue act structure of a meeting.

Each dialog act in a meeting is given one of 15 labels, which fall into six major groups:

- 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: a remainder class for utterances which convey an intention, but do not fit into the four previous categories;
- Backchannel, Stall and Fragment: classes for utterances without content, which allow complete segmentation of the material;

We have addressed the tasks of automatically segmenting the speech into dialogue acts, and assigning a label to each segment. The segmentation problem is non-trivial, since a single stretch of speech (with no pauses) from a speaker may comprise several dialogue acts—and conversely a single dialogue act may contain pauses.

Our approach to dialogue act recognition is based on a switching dynamic Bayesian network architecture which models a set of features related to lexical content and prosody and incorporates a weighted interpolated factored language model [29]. The switching DBN coordinates the recognition process by integrating all the available resources. The factored language model, which is learned from multiple conversational data corpora, is used in conjunction with additional task specific language models. In conjunction with this joint generative model, we have also employed a discriminative approach, based on conditional random fields, to perform a reclassification of the segmented DAs.

We have performed experiments using both automatic and manual transcriptions. The degradation when moving from manual transcriptions to the output of a speech recogniser is less than 10% absolute for both dialogue act classification and segmentation. Our experiments indicate that it is possible to perform automatic segmentation into DA units with a relatively low error rate. However the operations of tagging and recognition into fifteen imbalanced DA categories have a relatively high error rate, even after discriminative reclassification, indicating that this remains a challenging task.

5.2 Topic segmentation

Structuring a lengthy meeting by topic (and sub-topic) is a useful way of navigating a recorded meeting. Similar to dialogue act recognition, the aim is to infer automatically the sequential structure of the meeting; it differs in that the fundamental units (topics) are typically many minutes in duration.

Following Galley et al [34], we have explored two basic approaches to this task [30]. An unsupervised approach, LCSeg, does not require a training set of hand-marked topic boundaries, but can automatically infer topic boundaries as points where the statistics of text change significantly. An alternative supervised approach learns the 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. In addition to locating topic segments, we have developed approaches to automatically generating labels for topics, based on the statistics of the automatically transcribed words that make up a topic.

If suitable training data is available (such as the AMI corpus), then it is possible to construct accurate topic segmentation systems using classifiers such as decision trees or conditional random fields. Both topic segmentation and topic labelling are relatively robust to speech recognition, with only small degradation in performance when comparing speech recognition output to hand transcriptions.

5.3 Summarization

The automatic generation of summaries provides a natural way to succinctly describe the content of a meeting, and is a very natural way for users to obtain information. In AMI 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, as described in [32]. It thus requires (as a minimum) the automatic speech transcription and 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 these features are important to develop accurate methods for extractive summarization. Furthermore we have 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 themselves condensed, by removing irrelevant portions [35]. Again, taking account of speech features such as the overall intonation contour of the dialogue act helps to improve the overall performance.

6 Evaluation

We have performed evaluation both at the component technology level and at the system level, and the AMI corpus was designed to support evaluation at both levels. At the component level, in addition to internal evaluations in a common setting, we have participated in—and contributed data to—the the NIST Meeting Recognition (RT) evaluations[4] and the CLEAR evaluations[5] of focus of attention and face detection. Additionally, the AMI corpus, together with automatic speech recognition output, was provided to the Cross Language

[4] http://www.nist.gov/speech/tests/rt/

[5] http://www.clear-evaluation.org/

Evaluation Forum[6] (CLEF) for their 2007 evaluation on cross-lingual question answering.

Collaborative evaluation protocols are under development for a number of areas including dominance relations, speech summarization, dialogue act segmentation and tagging. These tasks are harder to evaluate compared with recognition tasks with an unambiguous ground truth, and there are several research challenges to address in developing these evaluations, relating to high inter-annotator disagreement, and the need for subjective human judgements.

Content extraction tasks, such as summarization or topic segmentation, are somewhat artificial as a standalone task, and are often carried out within some other context (such as browsing). In such cases, *extrinsic evaluation* approaches may be preferred, in which a task is evaluated in the context of a larger scenario, such as a meeting browser. In AMI we have developed a framework for extrinsic evaluation of browser components, that we call the *Browser Evaluation Test* (BET) [36]. 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).

More recently, we have developed a task-based evaluation [37] that is supported by the design of the AMI corpus. As outlined above, about 70% of corpus meetings are based on a replicable design team scenario. In the current version of 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.

7 Conclusions and future work

We have provided an overview of our work on the AMI and AMIDA projects. The major achievements of AMI include: the development of an instrumented meeting room infrastructure; the collection, annotation and release of the AMI meeting corpus; the development of a number of audio-video recognition technologies, in particular speech recognition for multi-party meetings; the development of multimodal content extraction approaches; and the development of novel frameworks for system-level 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. In our current work, we are paying particular attention to the integration of our existing recognition and content extraction modules into a framework of "meeting assistants" that can perform in close-to real-time (i.e., in some cases delays of several seconds or even minutes may be acceptable). We are interested in building applications that integrate these techniques for use during, and between, meetings in remote and co-located settings.

References

[1] R. Kazman, R. Al Halimi, William Hunt, and Marilyn Mantei, "Four paradigms for indexing video conferences," *IEEE Multimedia*, vol. 3, no. 1, 1996.

[2] D. M. Roy and S. Luz, "Audio meeting history tool: Interactive graphical user-support for virtual audio meetings," in *Proc. ESCA Workshop on Accessing Information in Spoken Audio*, 1999, pp. 107–110.

[3] R. Yong, A. Gupta, and J. Cadiz, "Viewing meetings captured by an omni-directional camera," *ACM Transactions on Computing Human Interaction*, March 2001.

[6] http://www.clef-campaign.org/

[4] D. Lee, B. Erol, and J. Graham, "Portable meeting recorder," *ACM Multimedia*, December 2002.

[5] 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 ICASSP*, May 2001.

[6] N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "Meetings about meetings: research at ICSI on speech in multiparty conversations," *in Proc. IEEE ICASSP*, 2003.

[7] L. Chen, R.T. Rose, F. Parrill, X. Han, J. Tu, Z. Huang, M. Harper, F. Quek, D. McNeill, R. Tuttle, et al., "VACE multimodal meeting corpus," *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2006.

[8] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation Journal*, 2007, In press.

[9] J. Carletta, S. Evert, U. Heid, and J. Kilgour, "The NITE XML toolkit: data model and query," *Language Resources and Evaluation Journal*, vol. 39, no. 4, pp. 313–334, 2005.

[10] D.A. van Leeuwen and M.A.H. Huijbregts, "The ami speaker diarization system for nist rt06s meeting data.," in *Proc. NIST RT06 Meeting Recognition Evaluation*, vol. 4299 of *Lecture Notes in Computer Science*, pp. 371–384. Springer Verlag, 2007.

[11] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audio-visual probabilistic tracking of multiple speakers in meetings," *IEEE Trans. on Audio Speech and Language Processing*, 2007.

[12] S. O. Ba and J.-M. Odobez, "A study on visual focus of attention recognition from head pose in a meeting room," in *Proc. MLMI '06*, 2006, AMI 176.

[13] A. Stolcke, C. Wooters, N. Mirghafori, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin, and M. Ostendorf, "Progress in meeting recognition: The ICSI-SRI-UW spring 2004 evaluation system," in *Proc. NIST RT04S Workshop.*, 2004.

[14] 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," in *Proceedings. ICASSP '07*, 2007.

[15] 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.

[16] J.S. Garofolo, C.D. Laprun, M. Miche, V.M. Stanford, and E. Tabassi, "The nist meeting room pilot corpus," in *Proc. 4th Intl. Conf. on Language Resources and Evaluation*, 2004.

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

[18] 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 *Proc. Interspeech'05*, 2005.

[19] J. Dines, J. Vepa, and T. Hain, "The segmentation of multi-channel meeting recordings for automatic speech recognition," in *Proc. Interspeech 2006*, 2006.

[20] J. Dines and J. Vepa, "Direct optimisation of a multilayer perceptron for the estimation of cepstral mean and variance statistics," in *Proc. Interspeech '07*, Antwerp, Belgium, 2007.

[21] 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. NIST RT'05 Workshop*, Edinburgh, 2005.

[22] P. Schwarz, P. Matějka, and J. Cernocký, "Towards lower error rates in phoneme recognition," in *Proc. of 7th Intl. Conf. on Text, Speech and Dialogue*, Brno, 2004, number ISBN 3-540-23049-1 in Springer, p. 8.

[23] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals, "The development of the AMI system for the transcription of speech in meetings," in *Proc. MLMI'05*, 2005.

[24] D. Povey, *Discriminative Training for Large Vocabulary Speech, Recognition*, Ph.D. thesis, Cambridge University, July 2004.

[25] D. Povey, M. J. F. Gales, D. Y. Kim, and P. C. Woodland, "MMI-MAP and MPE-MAP for acoustic model adaptation," in *Proc. Eurospeech'03*, 2003.

[26] M. Karafiat, L. Burget, J. Cernocky, and T. Hain, "Application of CMLLR in narrow band wide band adapted systems," in *Proc. Interspeech 2007*, Antwerp, Belgium, 2007.

[27] V. Wan and T. Hain, "Strategies for language model web-data collection," in *Proc. ICASSP'06*, 2006.

[28] T. Hain, L. Burget, J. Dines, M. Karafiat, D. van Leeuwen, M. Lincoln, G. Garau, and V. Wan, "The 2007 AMI(DA) system for meeting transcription," in *Proc. NIST RT07 Workshop*, 2007.

[29] Alfred Dielmann and Steve Renals, "DBN based joint dialogue act recognition of multiparty meetings," in *Proc IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP '07)*, 2007.

[30] P.-Y. Hsueh and J. Moore, "Automatic topic segmentation and labeling in multiparty dialogue," in *Proc IEEE/ACL SLT '06*, 2006, AMI-203.

[31] R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post, "Detection and application of influence rankings in small group meetings," *Proc ICMI '06*, 2006, AMI-192.

[32] G. Murray, S. Renals, J. Moore, and J. Carletta, "Incorporating speaker and discourse features into speech summarization," in *Proceedings of the Human Language Technology Conference of the NAACL*, 2006, pp. 367–374.

[33] M. Al-Hames, B. Hörnler, C. Scheuermann, and G. Rigoll, "Using audio, visual, and lexical features in a multi-modal virtual meeting director," in *Proc. MLMI '06*, 2006, AMI-164.

[34] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *Proc ACL '03*, 2003, pp. 21–26.

[35] G. Murray and S. Renals, "Dialogue act compression via pitch contour preservation," in *Proc. Interspeech '06*, 2006.

[36] P. Wellner, M. Flynn, S. Tucker, and S. Whittaker, "A meeting browser evaluation test," in *Proc. ACM CHI '05*, 2005, pp. 2021–2024.

[37] W. Post, M. A. A. in 't Veld, and S. A. A. van den Boogaard, "Evaluating meeting support tools," *Personal and Ubiquitous Computing*, 2007.