INTERPRETATION OF MULTIPARTY MEETINGS THE AMI AND AMIDA PROJECTS

Steve Renals (1), Thomas Hain (2) and Hervé Bourlard (3)

(1) Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH3 5EU, UK
(2) Dept. of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
(3) IDIAP Research Institute, 1920 Martigny, Switzerland
s.renals@ed.ac.uk t.hain@dcs.shef.ac.uk herve.bourlard@idiap.ch

ABSTRACT

The AMI and AMIDA projects are collaborative EU projects concerned with the automatic recognition and interpretation of multiparty meetings. This paper provides an overview of the advances we have made in these projects with a particular focus on the multimodal recording infrastructure, the publicly available AMI corpus of annotated meeting recordings, and the speech recognition framework that we have developed for this domain.

Index Terms— Meetings; speech recognition; AMI corpus; evaluation

1. INTRODUCTION

Since the mid-1990s a number of researchers have investigated the automatic recording, recognition and interpretation of meetings, e.g. [1, 2]. 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 automatic interpretation of human communication in meetings, with a particular emphasis on the development of approaches that help people to interact more effectively in meetings, and to easily access related information (including parts of previous meetings). The practical motivation for this research has come from two principal directions. First, the development of "smart" instrumented meeting rooms which are able to recognize and track the content of a multiparty meetings. Second, the need to develop more effective interfaces for remote participation in meetings that can provide a similar sense of "presence" and engagement compared with face-to-face meetings.

Communication in multiparty meetings is multimodal, factored across modalities including speech, gesture, projected displays and handwritten notes. The automatic recording, recognition and interpretation of such *communication scenes* is a considerable scientific challenge: separating communication information from multiple sources, integrating information from source across multiple modalities, and relating "lower level" information extracted from the signals to "higher level" semantics. Within the AMI consortium we have addressed these problems through a general approach of building statistical models using annotated corpora. A multidisciplinary approach has been crucial with active collaboration between researchers from speech recognition, computer vision, machine learning, social and organizational psychology, computational linguistics and human-computer interaction.

In section 2 we discuss the capture and development of the multimodal AMI corpus of multiparty meetings, using an instrumented meeting room (recently extended to enable the capture of remotely connected participants). Based upon this corpus we have developed and evaluated a number of audio-video recognizers, outlined in section 3. A major focus of the project is the development of automatic speech transcription for multiparty meetings, from both close-talking microphones and microphone arrays, discussed in section 4. Based upon the outputs of the multimodal recognizers we have developed a number of automatic approaches to content extraction (section 5) and a number of prototype applications that are concerned both with online meeting support and efficient access to meeting archives.

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. These rooms, which were designed for the collection of four person meetings, all contained a set of standardized recording equipment that included: six cameras (four providing close-up views of the participants, two providing a view of the whole room); twelve microphones (a headset microphone per

This work is supported by the European IST Programme Project FP6-0033812 (AMIDA). This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

[1] `http://www.amiproject.org/`

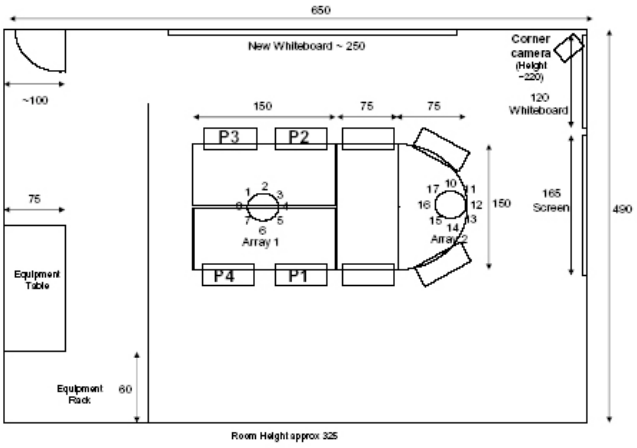

Fig. 1. Overhead schematic view of the instrumented meeting room at Edinburgh showing participant positions (P1–P4), two circular microphone arrays, and a room-view camera position (the second room-view camera was a ceiling-mounted overhead camera).

participant and an 8-element circular microphone array); data projector capture (VGA); whiteboard capture; and digital pen capture. There were also additional recording devices in each of the rooms, including an additional microphone array (Edinburgh), a binaural manikin (IDIAP) and additional cameras (IDIAP). A schematic plan of the instrumented meeting room at Edinburgh is shown in figure 1.

These instrumented meeting rooms were used to record the AMI Meeting Corpus [3], which consists of 100 hours of multimodal 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) [4], an open source XML-based infrastructure for the annotation and management of multimodal recordings[2].

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. It has already been employed for a number of international evaluations including the NIST Rich Transcription evaluations[3], the CLEAR evaluation[4] and the CLEF question-answering evaluation[5]. A number of "spoke" corpora have also been collected using the AMI instrumented meeting rooms, including the multi-channel Wall Street Journal audio-video (MC–WSJ–AV) corpus in which sentences from the Wall Street Journal speech recognition database were recorded (using lapel, headset and microphone arrays) in a variety of conditions including single stationary speaker, single moving speaker, and concurrent stationary speakers [5], which was used for the PASCAL Speech Separation Challenge 2 (SSC-2) [6, 7]

3. MULTIMODAL RECOGNITION

In the AMI project we have developed a number of recognizers for the multimodal meeting recordings, including speech transcription (discussed below), speaker diarization [8], audio-video localization and tracking [9], and visual focus of attention [10]. The outputs of these recognizers may be used directly, e.g. in a meeting browser, or as input for the automatic structuring or extraction of content from meetings.

4. MEETING SPEECH RECOGNITION

In this section we briefly describe the essential components of a meeting transcription system [11] and its accuracy in recent evaluations. The system is targeted on conference room meetings (as opposed to lectures or seminars), with the audio captured using both individual headset microphones (IHM) and microphone arrays (multiple distant microphones, MDM). The MDM is a more challenging speech recognition task, due to the additional reverberation and interference from other acoustic sources. We have employed a standard ASR framework using hidden Markov model (HMM) based acoustic modeling and N-gram based language models (LMs). Since an order of magnitude more transcribed data is available from domains such as conversational telephone speech (CTS), our system is bootstrapped from acoustic models trained on CTS (about 2000 hours in total), adapted using about 170 hours of multiparty meeting data from the AMI and ICSI [2] corpora.

Prior to recognition, the captured audio is pre-processed to address several issues including cross-talk detection and suppression for the IHM condition [12] and delay-sum beamforming and speaker segmentation and clustering for the MDM condition (fig. 2). 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).

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

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

[4] `http://www.clear-evaluation.org/`

[5] `http://www.clef-campaign.org/`

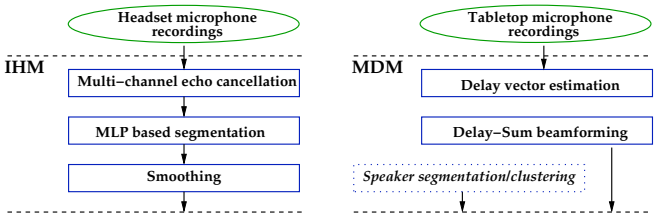

Fig. 2. Front-end processing stages for IHM and MDM.

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

The processing operates in several stages. First gain calibration is performed by normalizing 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 generalized 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. Although 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.

Twelve MF-PLP 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 [13] 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 normalization (VTLN) [14]. Secondly, heteroscedastic linear discriminant analysis (HLDA) gives consistent performance improvements [14]. Further gains can be obtained by discriminative training based on the minimum phone error (MPE) criterion, also jointly with constrained maximum likelihood regression (MLLR) based speaker adaptive training (SAT). The left column of Table 1

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 2. %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 3. %WER results on MDM data of the AMI 2007 system on the NIST RT'07 evaluation set.

shows WER results for models trained on 100 hours of meeting data and the *rt05seval* test set. In both cases substantial improvements are found.

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. Table 2 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 3 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%.

5. EXTRACTION OF STRUCTURE AND CONTENT

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, with an emphasis on models and algorithms that combine modalities.

Our work in this area has included the development of automatic approaches to the segmentation and classification of phenomena such as dialogue acts [15], topics [16], and dominance and influence [17], as well as abstractive and extractive summarization [18] and content-based automatic camera selection [19]. 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.

6. CONCLUSIONS

This paper has provided an overview of our work on the recognition and interpretation of multiparty meetings with a focus on speech recognition. In addition to the basic recognition and content extraction technologies described here, we have also developed a number of prototype applications including a number of configurable meeting browsers, and approaches to implicit extraction of information from meetings, based on the current context, such as the automatic linking of recorded previous meeting excerpts based on an online estimate of the content of a running meeting.

7. REFERENCES

[1] A. Waibel et al, "Advances in automatic meeting record creation and access," *in Proc. IEEE ICASSP*, May 2001.

[2] N. Morgan et al, "Meetings about meetings: research at ICSI on speech in multiparty conversations," *in Proc. IEEE ICASSP*, 2003.

[3] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation Journal*, vol. 41, pp. 181–190, 2007.

[4] 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.

[5] M. Lincoln, I. McCowan, J. Vepa, and H.K. Maganti, "The Multi-Channel Wall Street Journal Audio Visual corpus (MC-WSJ-AV): Specification and initial experiments," in *Proc. IEEE ASRU*, 2005, pp. 357–362.

[6] I. Himawan, I. McCowan, and M. Lincoln, "Microphone array beamforming approach to blind speech separation," in *Proc. MLMI'07 (LNCS 4892)*. 2008, pp. 295–305, Springer.

[7] J. McDonough et al, "To separate speech: A system for recognizing simultaneous speech," in *Proc. MLMI'07 (LNCS 4892)*. 2008, pp. 283–294, Springer.

[8] 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.

[9] 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.

[10] 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.

[11] T. Hain et al, "The AMI system for the transcription of speech in meetings," in *Proceedings. ICASSP '07*, 2007.

[12] 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.

[13] P. Schwarz, P. Matìjka, 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.

[14] T. Hain et al, "The development of the AMI system for the transcription of speech in meetings," in *Proc. MLMI'05*, 2005.

[15] A. Dielmann and S. Renals, "Recognition of dialogue acts in multiparty meetings using a switching DBN," *IEEE Trans. Audio, Speech and Language Processing*, 2008, in press.

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

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

[18] 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.

[19] 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.