Recognition and Understanding of Meetings

Recognition and Understanding of Meetings

Steve Renals
Centre for Speech Technology Research, University of Edinburgh
Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, UK
s.renals@ed.ac.uk homepages.inf.ed.ac.uk/srenals/

Abstract

This paper is about interpreting human communication in meetings using audio, video and other signals. Automatic meeting recognition and understanding is extremely challenging, since communication in a meeting is spontaneous and conversational, and involves multiple speakers and multiple modalities. This leads to a number of significant research problems in signal processing, in speech recognition, and in discourse interpretation, taking account of both individual and group behaviours. Addressing these problems requires an interdisciplinary effort. In this paper, I discuss the capture and annotation of multimodal meeting recordings—resulting in the AMI meeting corpus—and how we have built on this to develop techniques and applications for the recognition and interpretation of meetings.

1 Introduction

On the face of it, meetings do not seem to form a compelling research area. Although many people spend a substantial fraction of their time in meetings (e.g. the 1998 3M online survey at http://www.3m.com/meetingnetwork/), for most people they are not the most enjoyable aspect of their work. However, for all the time that is spent in meetings, technological support for the meeting process is scant. Meeting records usually take the form of brief minutes, personal notes, and more recent use of collaborative web 2.0 software. Such records are labour intensive to produce—because they are manually created—and usually fail to capture much of the content of a meeting, for example the factors that led to a particular decision and the different subjective attitudes displayed by the meeting attendees. For all the time invested in meetings, very little of the wealth of information that is exchanged is explicitly preserved.

To preserve the information recorded in meetings, it is necessary to capture it. Obviously this involves recording the speech of the meeting participants. However, human communication is a multimodal activity with information being exchanged via gestures, handwritten diagrams, and numerous social signals. The creation of a rich meeting record involves the capture of data across several modalities. It is a key engineering challenge to capture such multimodal signals in a reliable, unobtrusive and flexible way, but the greater challenges arise from unlocking the multimodal recordings. If such recordings are not transcribed and indexed (at the least), then access merely corresponds to replay. And it is rare that people will have the time, or the inclination, to replay a meeting.

There is a long and interesting thread of research which is concerned to better understand the dynamics of meetings and the way that groups function (Bales, 1951; McGrath, 1991; Stasser and Taylor, 1991). The types of analyses and studies carried out by these authors is still someway beyond what we can do automatically. The first significant work on automatic processing of meetings, coupled with an exploration of how people might interact with an archive of recorded meetings, was performed in the mid 1990s (Kazman et al., 1996). This work was limited by the fact that it was not possible at

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 1–9,
Los Angeles, California, June 2010.

that time to transcribe meeting speech automatically. Other early work in the area concentrated on the multimodal capture and broadcast of meetings (Roy and Luz, 1999; Cutler et al., 2002; Yong et al., 2001).

Three groups further developed approaches to automatically index the content of meetings. A team at Fuji Xerox PARC used video retrieval techniques such as keyframing to automatically generate manga-style summaries of meetings (Uchihashi et al., 1999), Waibel and colleagues at CMU used speech recognition and video tracking for meetings (Waibel et al., 2001), and Morgan and colleagues at ICSI focused on audio-only capture and speech recognition (Morgan et al., 2003). Since 2003 research in the recognition and understanding of meetings has developed substantially, stimulated by evaluation campaigns such as the NIST Rich Transcription (RT)[1] and CLEAR[2] evaluations, as well as some large multidisciplinary projects such as AMI/AMIDA[3], CHIL[4] and CALO[5].

This paper is about the work we have carried out in meeting capture, recognition and interpretation within the AMI and AMIDA projects since 2004. One of the principal outputs of these projects was a multimodal corpus of meeting recordings, annotated at a number of different levels. In section 2 we discuss collection of meeting data, and the construction of the AMI corpus. The remainder of the paper discusses the automatic recognition (section 3) and interpretation (section 4) of multimodal meeting recordings, application prototypes (section 5) and issues relating to evaluation (section 6).

2 The AMI corpus

Ideally it would not be necessary to undertake a large scale data collection and annotation exercise, every time we address a new domain. However unsupervised adaptation techniques are still rather immature, and prior to the collection of the AMI corpus, there had not been a controlled collection and multilevel annotation of multiparty interactions, recorded across multiple modalities.

[1] www.itl.nist.gov/iad/mig/tests/rt/
[2] clear-evaluation.org/
[3] www.amiproject.org/
[4] chil.server.de/
[5] caloproject.sri.com/

Figure 1: AMI instrumented meeting room: four co-located participants, one joined by video conference. In this case two microphone arrays and seven cameras were used.

One of our key motivations is the development of automatic approaches to recognise and interpret group interactions, using information spread across multiple modalities, but collected as unobtrusively as possible. This led to the design and construction of the AMI Instrumented Meeting Rooms (figure 1) at the University of Edinburgh, Idiap Research Institute, and TNO Human Factors. These rooms contained a set of standardised recording equipment including six or seven cameras (four of which would be used for close-up views in meeting of up to four people), an 8-element microphone array, a close-talking microphone for each participant (used to guarantee a clean audio signal for each speaker), as well capture of digital pens, whiteboards, shared laptop spaces, data projector and videoconferencing if used. A considerable amount of hardware was necessary for ensuring frame-level synchronisation. More recently we have used a lighter weight setup, that uses a high resolution spherical digital video camera system, and a single microphone array (7–20 elements, depending on meeting size) synchronised using software. We have also constructed a prototype system using a low-cost, flexible array of digital MEMS microphones (Zwyssig et al., 2010).

We used these instrumented meeting rooms to record the AMI Meeting Corpus (Carletta, 2007). This corpus contains over 100 hours of meeting recordings, with the different recording streams synchronised to a common timeline. The corpus contains a number of manually created and automatic annotations, synchronised to the same timeline. This

includes a high-quality manual word-level transcription of the complete corpus, as well as reference automatic speech recognition output, using the speech recognition system discussed in section 3 (using 5-fold cross-validation). In addition to word-level 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 head pose information. Some of these annotations, in particular video annotation, are expensive to perform: about 10 hours of meetings have been completely annotated at all these levels; over 70% of the corpus has been fully annotated with the linguistic annotations. NXT—the NITE XML Toolkit[6]—an XML-based open source software infrastructure for multimodal annotation was used to carry out and manage the annotations.

About 70% of the AMI corpus consists of meetings based on a design scenario, in which four participants play roles in a design team. The scenario involves four team meetings, between which the participants had tasks to accomplish. The participant roles were stimulated in real-time by email and web content. Although the use of a scenario reduces the overall realism of the meetings, we adopted this approach for several reasons, most importantly: (1) there were some preferred design outcomes, making it possible to define some objective group outcome measures; (2) the knowledge and motivation of the participants was controlled, thus removing the serious confounding factors that would arise from the long history and context found in real organisations; and (3) allowing the meeting scenario to be replicated, thus enabling system-level evaluations (as discussed in section 6). We recorded and annotated thirty replicates of the scenario: this provides an unparalleled resource for system evaluation, but also reduces the variability of the corpus (for example in terms of the language used). The remaining 30% of the corpus contains meetings that would have occurred anyway; these are meetings with a lot less control than the scenario meetings, but with greater linguistic variability.

All the meetings in the AMI corpus are spoken in English, but over half the participants are non-native speakers. This adds realism in a European context, as well as providing an additional speech recognition challenge. The corpus is publicly available[7], and is released under a licence that is based on the Creative Commons Attribution NonCommercial ShareAlike 2.5 Licence. This includes all the signals and manual annotations, plus a number of automatic annotations (e.g. speech recognition) made available to lower the startup cost of performing research on the corpus.

3 Multimodal recognition

The predominant motivation behind the collection and annotation of the AMI corpus was to enable the development of multimodal recognisers to address issues such as speech recognition, speaker diarisartion (Wooters and Huijbregts, 2007), gesture recognition (Al-Hames et al., 2007) and focus of attention (Ba and Odobez, 2008). Although speech recognition is based on the (multichannel) audio signal, the other problems can be successfully addressed by combining modalities. (There is certainly information in other modalities that has the potential to make speech recognition more accurate, but so far we have not been able to use it consistently and robustly.)

Speech recognition: The automatic transcription of what is spoken in a meeting is an essential prerequisite to interpreting a meeting. Morgan et al (2003) described the speech recognition of meetings as an "ASR-complete" problem. Developing an accurate system for meeting recognition involves the automatic segmentation of the recording into utterances from a single talker, robustness to reverberation and competing acoustic sources, handling overlapping talkers, exploitation of multiple microphone recordings, as well as the core acoustic and language modelling problems that arise when attempting to recognise spontaneous, conversational speech.

Our initial systems for meeting recognition used audio recorded with close-talking microphones, in order to develop the core acoustic modelling techniques. More recently our focus has been on recognising speech obtained using tabletop microphone

[6] sourceforge.net/projects/nite/

[7] corpus.amiproject.org/

arrays, which are less intrusive but have a lower signal-to-noise ratio. Multiple microphone systems are based on microphone array beamforming in which the individual microphone signals are filtered and summed to enhance signals coming from a particular direction, while suppressing signals from competing locations (Wölfel and McDonough, 2009).

The core acoustic and language modelling components for meeting speech recognition correspond quite closely to the state-of-the-art systems used in other domains. Acoustic modelling techniques include vocal tract length normalisation, speaker adaptation based on maximum likelihood linear transforms, and further training using a discriminative minimum Bayes risk criterion such as minimum phone error rate (Gales and Young, 2007; Renals and Hain, 2010). In addition we have employed a number of novel acoustic parameterisations including approaches based on local posterior probability estimation (Grezl et al., 2007) and pitch adaptive features (Garau and Renals, 2008), the automatic construction of domain-specific language models using documents obtained from the web by searching with n-grams obtained from meeting transcripts (Wan and Hain, 2006; Bulyko et al., 2007), and automatic approaches to acoustic segmentation optimised for meetings (Wrigley et al., 2005; Dines et al., 2006).

A feature of the systems developed for meeting recognition is the use of multiple recognition passes, cross-adaptation and model combination (Hain et al., 2007). In particular successive passes make use of more detailed—and more diverse—acoustic and language models. Different acoustic models trained on different feature representations (e.g. standard PLP features and posterior probability-based features) are cross-adapted, and different feature representations are also combined using linear transforms such as heteroscedastic linear discriminant analysis (Kumar and Andreou, 1998).

These systems have been evaluated in successive NIST RT evaluations: the core microphone array based system has a word error rate of about 40%; after adaptation and feature combination steps, this error rate can be reduced to about 30%. The equivalent close-talking microphone system has baseline word error rate of about 35%, reduced to less than 25% after further recognition passes (Hain et al., 2007). The core system runs about five times slower than real-time, and the full system is about fourteen times slower than real-time, on current commodity hardware. We have developed a low-latency real-time system (with an error rate of about 41% for microphone array input) (Garner et al., 2009), based on an open source runtime system[8].

4 Meeting interpretation

One of the interdisciplinary joys of working on meetings is that researchers with different approaches are able to build collaborations through working on common problems and common data. The automatic interpretation of meetings is a very good example: meetings form an exciting challenge for work in things such as topic identification, summarisation, dialogue act recognition and the recognition of subjective content. Although text-based approaches (using the output of a speech recognition system) form strong baselines, it is often the case that systems can be improved through the incorporation of information characteristic of spoken communication, such as prosody and speaker turn patterns, as well video information such as head or hand movements.

Segmentation: We have explored multistream statistical models to automatically segment meeting recordings. Meetings can be usefully segmented at many different levels, for example into speech and non-speech (an essential pre-processing for speech recognition), into utterances spoken by a single talker, into dialogue acts, into topics, and into "meeting phases". The latter was the subject of our first investigations in using multimodal multistream models to segment meetings.

Meetings are group events, characterised by both individual actions and group actions. To obtain structure at the group level, we and colleagues in the M4 and AMI projects investigated segmenting a meeting into a sequence of group actions such as monologue, discussion and presentation (McCowan et al., 2005). We used a number of feature streams for this segmentation and labelling task including speaker turn dynamics, prosody, lexical information,

[8] juicer.amiproject.org/

and participant head and hand movements (Dielmann and Renals, 2007). Our initial experiments used an HMM to model the feature streams with a single hidden state space, and resulted in an "action error rate" of over 40% (action error rate is analogous to word error rate, but defined over meeting actions, presumed not to overlap). The HMM was then substituted by a richer DBN multistream model in which each feature stream was processed independently at a lower level of the model. These partial results were then combined at a higher level, thus providing hierarchical integration of the multimodal feature streams. This multistream approach enabled a later integration of feature streams and increased flexibility in modelling the interdependences between the different streams, enabling some accommodation for asynchrony and multiple time scales. Thus use of the richer DBN multistream model resulted in a significant lowering of the action error rate to around 13%.

We extended this approach to look at a much finer grained segmentation: dialogue acts. A dialogue act can be viewed as a segment of speech labelled so as to roughly categorise the speaker's intention. In the AMI corpus each dialogue act in a meeting is given one of 15 labels, which may be categorised as information exchange, making or eliciting suggestions or offers, commenting on the discussion, social acts, backchannels, or "other". 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. To address the tasks of automatically segmenting the speech into dialogue acts, and assigning a label to each segment, we employed a switching dynamic Bayesian network architecture, which modelled a set of features related to lexical content and prosody and incorporates a weighted interpolated factored language model (Dielmann and Renals, 2008). The switching DBN coordinated the recognition process by integrating all the available resources. This approach was able to leverage additional corpora of conversational data by using them as training data for a factored language model which was used in conjunction with additional task specific language models. We followed this joint generative model, with a discriminative approach, based on conditional random fields, which performed a reclassification of the segmented dialogue acts.

Our experiments on dialogue act recognition used both automatic and manual transcriptions of the AMI corpus. The degradation when moving from manual transcriptions to the output of a speech recogniser was less than 10% absolute for both dialogue act classification and segmentation. Our experiments indicated that it is possible to perform automatic segmentation into dialogue acts 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.

Summarisation: The automatic generation of summaries provides a natural way to succinctly describe the content of a meeting, and can be an efficient way for users to obtain information. We have focussed on extractive techniques to construct summaries, in which the most relevant parts of a meeting are located, and concatenated together to provide a 'cut-and-paste' summary, which may be textual or multimodal.

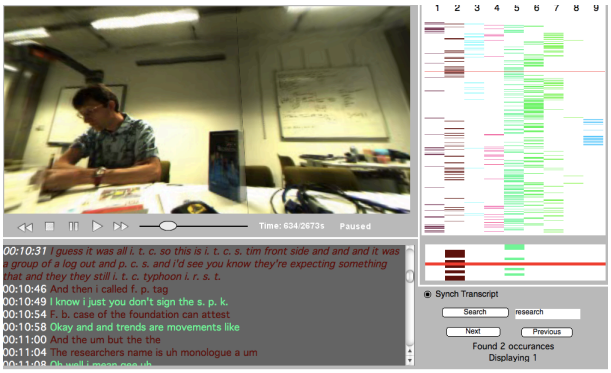
Our approach to extractive summarisation is based on automatically extracting relevant dialogue acts (or alternatively "spurts", segments spoken by a single speaker and delimited by silence) from a meeting (Murray et al., 2006). This requires (as a minimum) the automatic speech transcription and, if spurts are not used, dialogue act segmentation. Lexical information is clearly extremely important for summarisation, but we have also found 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) to be important for the development of accurate methods for extractive summarisation of meetings. Furthermore we have explored reduced dimension representations of text, based on latent semantic analysis, which we found added precision to the summarisation. Using an evaluation measure referred to as weighted precision, we discovered that it is possible to reliably extract the most relevant dialogue acts, even in the presence of speech recognition errors.

5 Application prototypes

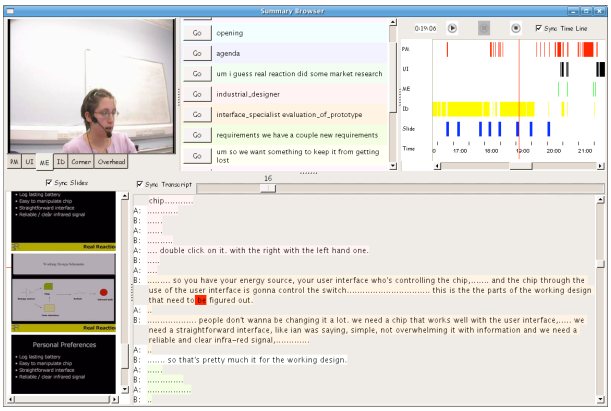
We have incorporated these meeting recognition and interpretation components in a number of applications. Our basic approach to navigating meeting archives centres on the notion of meeting browsers, in which media files, transcripts and segmentations are synchronised to a common time line. Figure 2 (a) gives an example of such a browser, which also enables a user to pan and zoom within the captured spherical video stream.

We have explored (and, as discussed below, evaluated) a number of ways of including automatically generated summaries in meeting browsers. The browser illustrated in figure 2 (b) enables navigation by the summarised transcript or via the topic segmentation. In this case the degree of summarisation is controlled by a slider, which removes those speech segments that do no contribute to the summary. We have also explored real-time (with a few utterances latency) approaches to summarisation, to facilitate meeting "catchup" scenarios, including the generation of audio only summaries, with about 60% of the speech removed (Tucker et al., 2010). Visualisations of summaries include a comic book layout (Castronovo et al., 2008), illustrated in figure 3. This is related to "VideoManga" (Uchihashi et al., 1999), but driven by transcribed speech rather than visually identified keyframes.

The availability of real-time meeting speech recognition, with phrase-level latency (Garner et al., 2009), enables a new class of applications. Within AMIDA we developed a software architecture referred to as "The Hub" to support real-time applications. The Hub is essentially a real-time annotation server, mediating between annotation producers, such as speech recognition, and annotation consumers, such as a real-time catchup browser. Of course many applications will be both producers and consumers: for instance topic segmentation consumes transcripts and speaker turn information and produces time aligned topic segments. A good example of an application made possible by real-time recognition components and the Hub is the AMIDA Content Linking Device (Popescu-Belis et al., 2008). Content linking is essentially a continual real-time search in which a repository is searched using a query constructed from the current conversational context. In this case the context is obtained from a speech recognition transcript of the past 30 seconds of the conversation, and a query is constructed using $tf \cdot idf$ or a similar measure, combined with predefined keywords or topic weightings. The repository to be searched may be the web, or a portion of the web, or it may be an organisational document repository, including transcribed, structured and indexed recordings of previous meetings. Figure 4 shows a basic interface to content linking. We have constructed live content-linking systems, driven by microphone array based real-time speech recognition, with the aim of presenting—without explicit query—potentially relevant documents to meeting participants.



(a) Basic web-based browser



(b) Summary browser

Figure 2: Two examples of meeting browsers, both include time synchronisation with a searchable ASR transcript and speaker activities. (a) is a basic web-based browser; (b) also employs extractive summarisation and topic segmentation components.

Figure 3: Comic book display of automatically generated meeting summary.

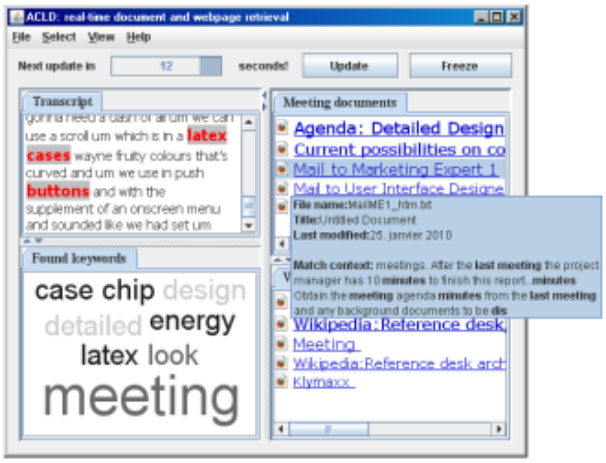


Figure 4: Demonstration screenshot of the AMI automatic content linking device. The subpanels show (clockwise from top left) the ASR transcript, relevant documents from the meeting document base, relevant web hits and a a tag cloud.

6 Evaluation

The multiple streams of data and multiple layers of annotations that make up the AMI corpus enable it to be used for evaluations of specific recognition components. The corpus has been used to evaluate many different things including voice activity detection, speaker diarisation and speech recognition (in the NIST RT evaluations), and head pose recognition (in the CLEAR evaluation). In the spoken language processing domain, the AMI corpus has been used to evaluate meeting summarisation, topic segmentation, dialogue act recognition and cross-language retrieval.

In addition to intrinsic component-level evaluations, it is valuable to evaluate complete systems, and components in a system context. In the AMI/AMIDA projects, we investigated a number of extrinsic evaluation frameworks for browsing and accessing meeting archives. The Browser Evaluation Test (BET) (Wellner et al., 2005) provides a framework for the comparison of arbitrary meeting browser setups, which may differ in terms of which content extraction or abstraction components are employed. In the BET test subjects have to answer true/false questions about a number of "observations of interest" relating to a recorded meeting, using the browser under test with a specified time limit (typically half the meeting length).

We developed of a variant of the BET to specifically evaluate different summarisation approaches. In the Decision Audit evaluation (Murray et al., 2009) the user's task is to ascertain the factors across a number of meetings that lead to a particular decision being made. A set of browsers were constructed differing in the summarisation approach employed (manual vs. ASR transcripts; extractive vs. abstractive vs. human vs. keyword-based summarisation), and the test subjects used them to perform the decision audit. Like the BET this evaluation is labour-intensive, but the results can be analysed using a battery of objective and subjective measures. Conclusions from carrying out this evaluation indicated that the task itself was quite challenging for users (even with human transcripts and summaries, most users could not find many factors involved in the decision), that automatic extractive summaries outperformed reasonably competitive baseline approaches, and that although subjects reported ASR transcripts to be unsatisfactory (due to the error rate) browsing using the ASR transcript still resulted in users' being generally able to find the relevant parts of the meeting archive.

7 Conclusions

In this paper I have given an overview of our investigations into automatic meeting recognition and interpretation. Multiparty communication is a challenging problem at many levels, from signal processing to discourse modelling. A major part of our attempt to address this problem, in an interdisciplinary way, was the collection, annotation, and distribution of the AMI meeting corpus. The AMI corpus has been at the basis of nearly all the work that we have carried out in the area, from speech recognition to summarisation. Multiparty speech recognition remains a difficult task, with a typical error rate of over 20%, however the accuracy is enough to enable various components to build on top of it. A major achievement has been the development of prototype applications that can use phrase-level latency real-time speech recognition.

Many of the automatic approaches to meeting recognition and characterisation are characterised by extensive combination at the feature stream, model and system level. In our experience, such approaches offer consistent improvements in accuracy for these complex, multimodal tasks.

Meetings serve a social function, and much of this has been ignored in our work, so far. We have focussed principally on understanding meetings in terms of their lexical content, augmented by various multimodal streams. However in many interactions, the social signals are at least as important as the propositional content of the words (Pentland, 2008); it is a major challenge to develop meeting interpretation components that can infer and take advantage of such social cues. We have made initial attempts to do this, by attempting to include aspects such as social role (Huang and Renals, 2008).

The AMI corpus involved a substantial effort from many individuals, and provides an invaluable resource. However, we do not wish to do this again, even if we are dealing with a domain that is significantly different, such as larger groups, or family "meetings". However, our recognisers rely strongly on annotated in-domain data. It is a major challenge to develop algorithms that are unsupervised and adaptive to free us from the need to collect and annotate large amount of data each time we are interested in a new domain.

Acknowledgments

This paper has arisen from a collaboration involving several laboratories. I have benefitted, in particular, from long-term collaborations with Hervé Bourlard, Jean Carletta, Thomas Hain, and Mike Lincoln, and from a number of fantastic PhD students. This work was supported by the European IST/ICT Programme Projects IST-2001-34485 (M4), FP6-506811 (AMI), FP6-033812 (AMIDA), and FP7-231287 (SSPNet). This paper only reflects the author's views and funding agencies are not liable for any use that may be made of the information contained herein.

References

M. Al-Hames, C. Lenz, S. Reiter, J. Schenk, F. Wallhoff, and G. Rigoll. 2007. Robust multi-modal group action recognition in meetings from disturbed videos with the asynchronous hidden Markov model. In *Proc IEEE ICIP*.

S. O. Ba and J. M. Odobez. 2008. Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *Proc. IEEE ICASSP*.

R. F. Bales. 1951. *Interaction Process Analysis*. Addison Wesley, Cambridge MA, USA.

I. Bulyko, M. Ostendorf, M. Siu, T. Ng, A. Stolcke, and O. Cetin. 2007. Web resources for language modeling in conversational speech recognition. *ACM Transactions on Speech and Language Processing*, 5(1):1–25.

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

S. Castronovo, J. Frey, and P. Poller. 2008. A generic layout-tool for summaries of meetings in a constraint-based approach. In *Machine Learning for Multimodal Interaction (Proc. MLMI '08)*. Springer.

R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg. 2002. Distributed meetings: a meeting capture and broadcasting system. In *Proc. ACM Multimedia*, pages 503–512.

A. Dielmann and S. Renals. 2007. Automatic meeting segmentation using dynamic Bayesian networks. *IEEE Transactions on Multimedia*, 9(1):25–36.

A. Dielmann and S. Renals. 2008. Recognition of dialogue acts in multiparty meetings using a switching DBN. *IEEE Transactions on Audio, Speech and Language Processing*, 16(7):1303–1314.

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

M. J. F. Gales and S. J. Young. 2007. The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304.

G. Garau and S. Renals. 2008. Combining spectral representations for large vocabulary continuous speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 16(3):508–518.

P. Garner, J. Dines, T. Hain, A. El Hannani, M. Karafiat, D. Korchagin, M. Lincoln, V. Wan, and L. Zhang. 2009. Real-time ASR from meetings. In *Proc. Interspeech*.

F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky. 2007. Probabilistic and bottle-neck features for lvcsr of meetings. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–757–IV–760.

T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan. 2007. The ami system for the transcription of speech in meetings. In *Proc. IEEE ICASSP–07*.

S. Huang and S. Renals. 2008. Unsupervised language model adaptation based on topic and role information in multiparty meetings. In *Proc. Interspeech '08*.

R. Kazman, R. Al-Halimi, W. Hunt, and M. Mantei. 1996. Four paradigms for indexing video conferences. *Multimedia, IEEE*, 3(1):63–73.

N. Kumar and A. G. Andreou. 1998. Heteroscedastic discriminant analysis and reduced rank HMMs for improved recognition. *Speech Communication*, 26:283–297.

I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. 2005. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):305–317.

J. E. McGrath. 1991. Time, interaction, and performance (TIP): A theory of groups. *Small Group Research*, 22(2):147.

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. 2003. Meetings about meetings: research at ICSI on speech in multiparty conversations. In *Proc. IEEE ICASSP*.

G. Murray, S. Renals, J. Moore, and J. Carletta. 2006. Incorporating speaker and discourse features into speech summarization. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 367–374.

G. Murray, T. Kleinbauer, P. Poller, T. Becker, S. Renals, and J. Kilgour. 2009. Extrinsic summarization evaluation: A decision audit task. *ACM Transactions on Speech and Language Processing*, 6(2):1–29.

A.S. Pentland. 2008. *Honest signals: how they shape our world*. The MIT Press.

A. Popescu-Belis, E. Boertjes, J. Kilgour, P. Poller, S. Castronovo, T. Wilson, A. Jaimes, and J. Carletta. 2008. The amida automatic content linking device: Just-in-time document retrieval in meetings. In *Machine Learning for Multimodal Interaction (Proc. MLMI '08)*.

S. Renals and T. Hain. 2010. Speech recognition. In A. Clark, C. Fox, and S. Lappin, editors, *Handbook of Computational Linguistics and Natural Language Processing*. Wiley Blackwell.

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

G. Stasser and LA Taylor. 1991. Speaking turns in face-to-face discussions. *Journal of Personality and Social Psychology*, 60(5):675–684.

S. Tucker, O. Bergman, A. Ramamoorthy, and S. Whittaker. 2010. Catchup: a useful application of time-travel in meetings. In *Proc. ACM CSCW*, pages 99–102.

S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky. 1999. Video manga: generating semantically meaningful video summaries. In *Proc. ACM Multimedia*, pages 383–392.

A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner. 2001. Advances in automatic meeting record creation and access. In *Proc IEEE ICASSP*.

V. Wan and T. Hain. 2006. Strategies for language model web-data collection. In *Proc IEEE ICASSP*.

P. Wellner, M. Flynn, S. Tucker, and S. Whittaker. 2005. A meeting browser evaluation test. In *Proc. ACM CHI*, pages 2021–2024.

M. Wölfel and J. McDonough. 2009. *Distant Speech Recognition*. Wiley.

C. Wooters and M. Huijbregts. 2007. The ICSI RT07s speaker diarization system. In *Multimodal Technologies for Perception of Humans. International Evaluation Workshops CLEAR 2007 and RT 2007*, volume 4625 of *LNCS*, pages 509–519. Springer.

S. Wrigley, G. Brown, V. Wan, and S. Renals. 2005. Speech and crosstalk detection in multichannel audio. *IEEE Transactions on Speech and Audio Processing*, 13(1):84–91.

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

E. Zwyssig, M. Lincoln, and S. Renals. 2010. A digital microphone array for distant speech recognition. In *Proc. IEEE ICASSP–10*.