Automatic analysis of multiparty meetings

Steve Renals
The Centre for Speech Technology Research
University of Edinburgh
UK

Abstract

This paper is about the recognition and interpretation of multiparty meetings captured as audio, video and other signals. This is a challenging task since the meetings consist of spontaneous and conversational interactions between a number of participants: it is a *multimodal*, *multiparty*, *multistream* problem. We discuss the capture and annotation of the AMI meeting corpus, the development of a meeting speech recognition system, and systems for the automatic segmentation, summarisation and social processing of meetings, together with some example applications based on these systems.

Keywords: speech recognition, multimodal interaction, AMI corpus, summarisation, segmentation, multiparty meetings.

1 INTRODUCTION

In this paper we are concerned with multiparty conversations, with a focus on *meetings*. Meetings contain realistic, complex behaviours in a circumscribed setting, and present a set of significant scientific challenges. They also form the basis of a number of potentially important applications.

Meetings have been collected for research since the 1940s (Bales 1951), and there is a large body of research in social psychology concerning the dynamics of group discussions and meetings, and the way that groups function (Stasser and Taylor 1991, McGrath 1991). Much of this work has been based on laborious hand annotation of meetings, often carried out in real-time while a meeting is underway. Since the early 1990s it has become possible to capture and annotate meetings automatically, as described in this paper. However the richness and sophistication of the analyses carried out E hand is still well beyond what can be done automatically.

Probably the first automatic work in this area was carried out by Kazman et al. (1996) who developed a tool to automatically index videoconferences. The approach to indexing in this work was based on lexical chaining algorithms operating on meeting transcripts; however, the system was limited by the then inadequate performance of speech recognition systems operating on multiparty conversations. Other early work on meeting capture and analysis focused on capture and broadcasting (Uchihashi et al. 1999, Roy and Luz 1999, Yong et al. 2001, Lee et al. 2002, Cutler et al. 2002), with particular emphases on panoramic video and multimodal capture.

Although, Uchihashi et al. (1999) developed a novel approach to browsing such meetings based on video keyframes presented as a manga-style comic book, useful semantic search across meeting recordings relies on a transcription of what was said (as recognised by Kazman). Research teams at Carnegie Mellon University (Waibel et al. 2003) and the International Computer Science Institute (Morgan et al. 2003) were the first to focus on the problem of speech recognition in meetings, as a key enabler to meeting analysis. Since that time there have been a number of large scale projects focusing on meeting recognition and interpretation including AMI (Augmented Multiparty Interaction: http://www.amiproject.org), CHIL (Computers in the Human Interaction Loop: http://chil.server.de), and CALO (Cognitive Assistant that Learns and Organizes: http://caloproject.sri.com).

Meeting recognition and analysis presents a substantial set of interdisciplinary research problems, reflected in the above-mentioned projects. Meetings take place in a natural setting and their analysis must take account of several underlying characteristics:

- Meetings are *multiparty*, involving communication between several people. Although individual behaviours can certainly be identified, a meeting cannot be fully comprehended without taking account of group behaviours and the social roles adopted by meeting participants.
- Meetings are *multimodal*. Communication in a meeting involves more than a sequence of words, the information transmitted in a meeting is factored across several modalities including the rhythm and timing of participant's speech, communicative gestures, and foci of attention.
- Meeting processing should be *multistream*. Given the multimodal, multiparty nature of meeting communication, it follows that multiple, asynchronous streams of data must be processed and combined to enable their analysis.

In this paper we discuss some of the research carried out in the AMI project (and its follow-on project AMIDA—AMI with Distant Access). These large interdisciplinary European Union projects had as their scientific goal the automatic recognition and understanding of human communication in meetings. To achieve this goal a new corpus infrastructure was required; section 2 discusses design, collection and annotation of the AMI meetings corpus. In section 3 we discuss meeting recognition, with a particular focus on the construction of a speech recognition system for multiparty meetings recorded using microphone arrays. Section 4 discusses the automatic structuring and extraction of content from meeting recordings, including segmentation, summarisation and the extraction of socially-related information. We close with a discussion of how these meeting recognition and content technologies may be integrated in some exemplar applications.

2 DATA, ANNOTATION AND THE AMI CORPUS

Speech recognition research has been through a number of phases: until the mid-1990s virtually all research was focused on dictation, command and control applications, and human-computer spoken dialogue systems. These systems each involved a single human talker speaking to a computer system; the human speech to be recognised was rather controlled and often pre-planned. This is different to natural human conversation, which features two or more interlocutors and spontaneous, conversational speech characterised by much greater variability and by phenomena such as talker

overlap. Research on conversational speech recognition began with the Switchboard corpus of two-party telephone dialogues (Godfrey et al. 1992) and the Map Task corpus of two-party conversations of people cooperating to solve a particular task (Anderson et al. 1991). These corpora contained highly variable, conversational speech, but did not address the multiparty conversations which fill our everyday lives.

The first substantial effort to build a research infrastructure for multi-party meetings came with the work of Carnegie Mellon University (Waibel et al. 2003) and the International Computer Science Institute (ICSI) (Morgan et al. 2003), both of whom focused on speech recognition in meetings, primarily recorded using headset microphones. ICSI collected, transcribed and released a 75-hour corpus of multiparty meetings (Janin et al. 2003). The ICSI meeting corpus was recorded using high quality individual headset microphones worn by each participant, together with a number of tabletop microphones. This corpus was first large scale research corpus to feature overlapping speech and the non-linear conversational structure that makes meeting speech recognition such a challenging task. The data was collected from naturally occurring research group meetings, and was the most important training data for the the first meeting speech recognition systems.

Meetings are multimodal events, and the design and construction of the AMI corpus (Carletta et al. 2005) reflect this. The AMI corpus comprises about 100 hours collected in three instrumented meeting rooms, designed and standardised by the AMI project, in the UK (University of Edinburgh), the Netherlands (TNO Research Institute), and Switzerland (Idiap Research Institute). Each instrumented meeting room was designed with a four-party meeting in mind, and contained six or seven cameras (four participant close-up, two or three room view), an eight-element circular microphone array, close-talking and lapel microphones for each participant (to ensure a clean audio signal for each talker), as well as digital pens, smart whiteboards, shared computer work spaces, data projection, and videoconferencing all captured and time-synchronised. Since the AMI corpus was designed to be for research use, considerable effort was made to ensure synchronisation between all the media streams. Frame-level synchronisation was ensured using a hardware time-signal generator. Camera views from the AMI corpus are shown in fig 1.

The AMI corpus thus consists of 100 hours of meeting recordings, synchronised to a common timeline. The corpus has been heavily annotated by hand, primarily to provide ground truth for machine learning. In addition to high quality speech transcription, AMI corpus annotations include topic boundaries, dialogue act boundaries and labels, named entities, and summaries. A number of video annotations were also made including head and hand gestures, head pose and person location. The annotations were made and managed using the NITE XML Toolkit (NXT), an open source toolkit for the annotation of multimodal corpora (http://groups.inf.ed.ac.uk/nxt/) (Carletta et al. 2003). NXT provides a data model and a query language for annotated multimodal corpora, as well as a set of interface components that can be used to create interfaces for annotation and corpus viewing. Figure 2 illustrates an NXT interface for the annotation of dialogue acts.

All the meetings in the AMI corpus were conducted in English, but there was a wide variety of accents and dialects spoken. Over half the participants in the meetings were non-native speak-

Fig. 1. Six camera views from an AMI corpus meeting recording. In a four-person meeting, there is a close-up camera pointing at each seating position, plus two room-view cameras. Each participant was recorded using a headset and a lapel microphone, and there were two circular 8-element microphone arrays on the meeting table.

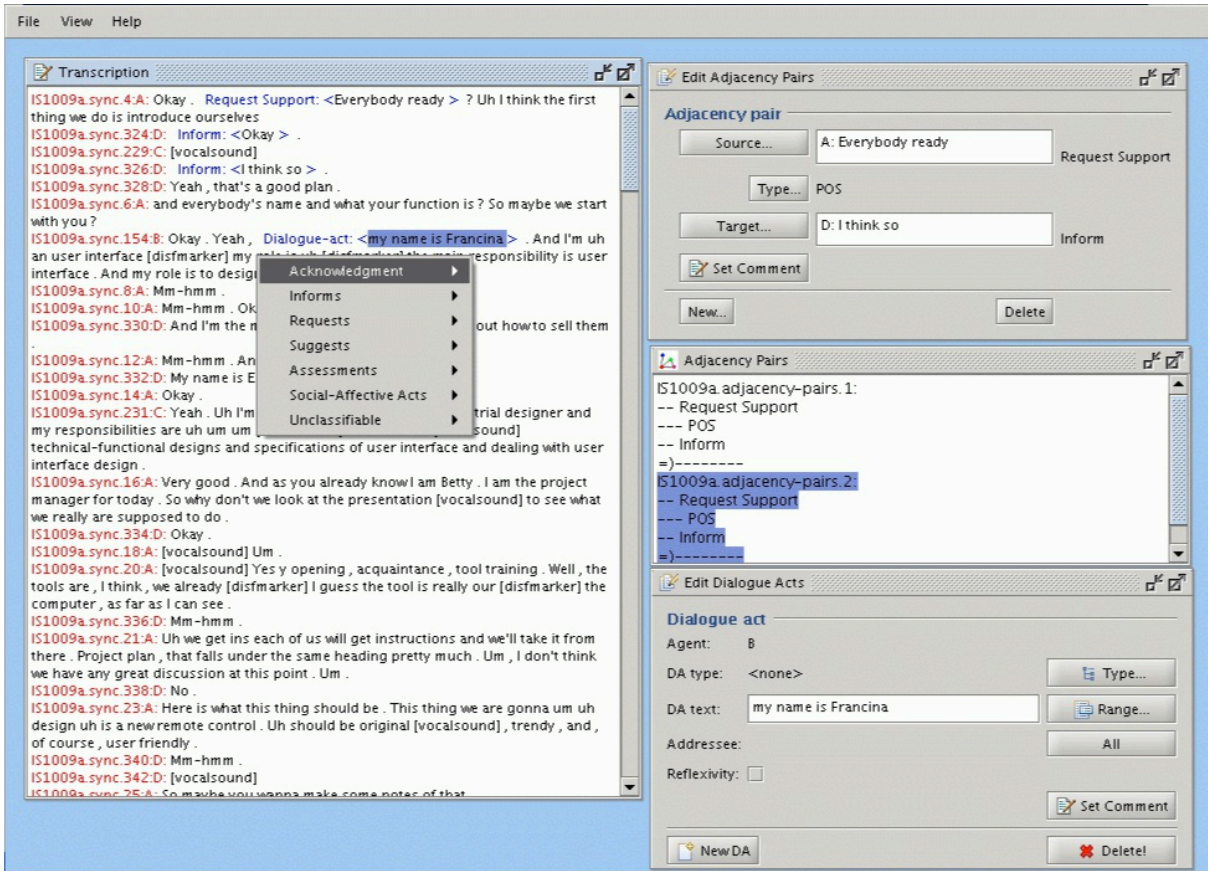

Fig. 2. NXT tool for annotating dialogue acts in a multiparty conversation.

ers of English. This adds realism, particularly in a European context. The corpus is freely available to researchers under a Creative Commons "ShareAlike Attribution NonCommercial" license (http://corpus.amiproject.org/). In addition to manual ground-truth annotations, the corpus also includes a number of baseline automatic annotations including transcripts produced by automatic speech recognition, topic boundaries and dialogue act segments. This allows researchers to compare the performance of their systems on both human and automatic annotations, without the need to construct, for example, an automatic speech recogniser.

In contrast to the ICSI meeting corpus, the AMI corpus includes both "scenario" meetings (about 70% of the corpus) and naturally occurring non-scenario meetings (the remaining 30%). The scenario meetings were based around a four-person team with the task of designing a remote control device for a TV. Each participant in the team was given a role (project manager, industrial designer, marketing expert, interface designer), and the team participated in a series of four meetings. The meetings took place over 3-4 hours: about half of this time was spent in meetings, the remainder was spent in preparation, with each participant having their role stimulated by real-time emails (e.g. budget information was mailed to the project manager) and specifically generated web content. Although such a scenario reduces the overall realism of the task, it brings a number of benefits:

1) **Control.** Scenario meetings enable the knowledge and motivation of participants to be controlled.

Unlike naturally occurring meetings, scenario teams do not have a (potentially long) history of previous interactions.

2) **Evaluation measures.** By defining a particular scenario, product design in the case of the AMI corpus, and controlling the knowledge of the participants, it is possible to define objective group outcome measures based on an optimal design outcome corresponding to the constraints defined in the scenario.
3) **Replicability.** Scenario meetings may be replicated multiple times, but in different experimental conditions. For example, the meetings may be carried out with one or more of the participants in a remote location, joining the meeting by videoconferencing.

In the AMI corpus we recorded thirty versions of the scenario, each with a different group. This enabled us to develop system-level evaluations (discussed further in section 4), at the cost of less diversity in the corpus.

3 MEETING RECOGNITION

One of the main motivations behind the capture and annotation of the AMI corpus was to enable the supervised training of multimodal recognition systems. Our aim was to develop recognisers to answer a number of key questions:

1) *Speech recognition*—What was said? (Hain et al. 2005, Garner et al. 2009)
2) *Localisation and tracking*—Who and where are the people in the meeting? (Gatica-Perez et al. 2007)
3) *Speaker diarisation*—Who spoke when? (Wooters and Huijbregts 2008)
4) *Gesture and action recognition*—How do people act? (Poppe and Poel 2008)
5) *Visual focus of attention*—Where or what are people attending to? (Ba and Odobez 2008)

A substantial amount of work has been carried out in each of these areas. In this paper we discuss meeting speech recognition.

Automatic speech transcription is fundamental to the overall recognition and interpretation of a meeting. Meeting speech recognition is extremely challenging: Morgan et al. (2003) refer to it as an "ASR Complete" problem. The construction of a system to transcribe speech in meetings involves the solution of many subproblems including speech/non-speech detection, robustness to reverberation, separation of multiple acoustic sources, exploitation of microphone arrays, detection and processing of overlapping talkers, as well as the well-known acoustic modelling and language modelling challenges that arise when dealing with spontaneous, conversational speech.

The AMI corpus included individual headset microphone (IHM) recordings in order to enable the initial development of acoustic models for meetings recognition, at first by maximum a-posteriori (MAP) adaptation of acoustic models for conversational telephone speech (Hain et al. 2005). However, the major focus has been on the use of multiple distant microphone (MDM) recordings as the main acoustic capture condition, using microphone array beamforming algorithms to filter and sum the individual microphone signals in order to enhance sounds from a particular direction, while suppressing other directions (Wölfel and McDonough 2009).

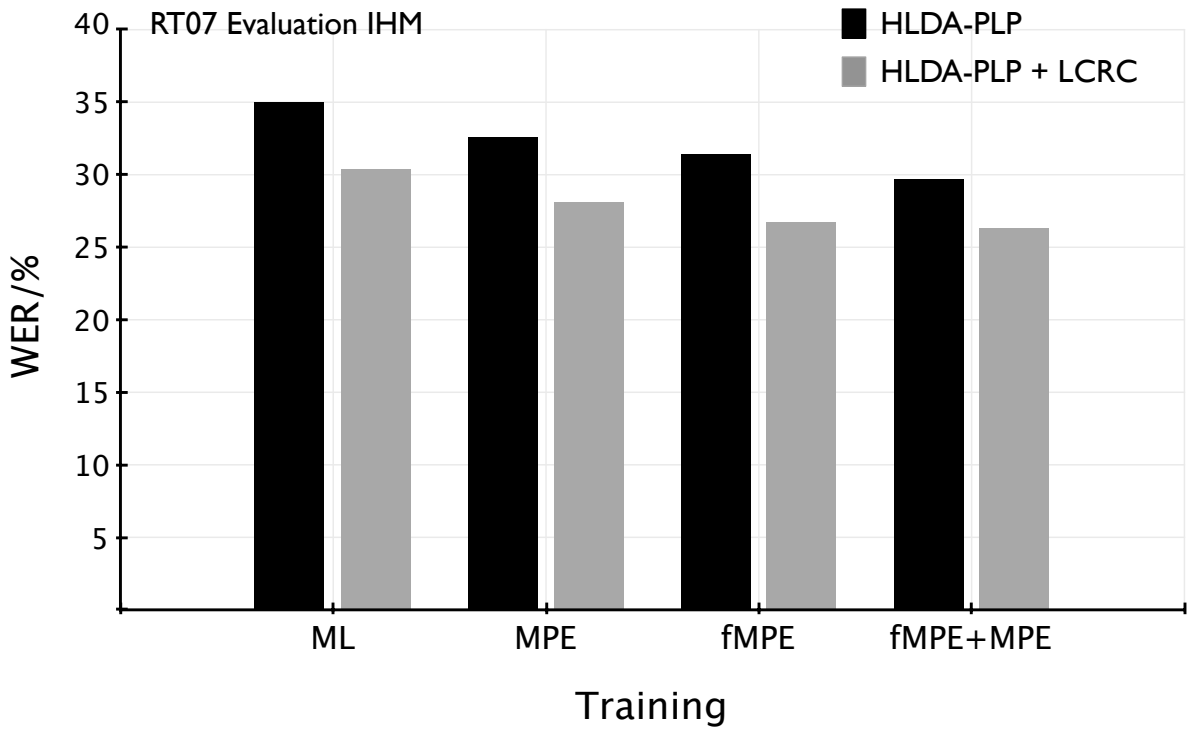

Fig. 3. Word error rates by the AMI-ASR system on the RT07 evaluation data from meeting recordings using individual headset microphones (IHM). The graph compares four training criteria (maximum likelihood — ML; minimum phone error rate — MPE; feature-based MPE — fMPE; and fMPE followed by MPE) and two feature sets (perceptual linear prediction (PLP) coefficients decorrelated by an HLDA transform — PLP-HLDA; and PLP-HLDA combined with long-context LCRC features — PLP-HLDA + LCRC).

The acoustic and language modelling components used for meeting speech recognition are similar to those used in state-of-the-art systems for domains such as conversational telephone speech or broadcast news (Chen et al. 2006, Gales et al. 2006). Such systems are based on context-dependent phone models, in which a phone in context is modelled by a hidden Markov model (HMM) with a Gaussian mixture model (GMM) probability density function. These acoustic models are usually combined with an n-gram language model to produce a complete system (Gales and Young 2007, Renals and Hain 2010). A system such as this, using mel frequency cepstral coefficient acoustic features, and trained using maximum likelihood estimation results in a word error rate of about 42% on meeting speech in the US National Institute of Standards and Technology Rich Transcription evaluation 2007 (NIST RT07; http://www.itl.nist.gov/iad/mig/tests/rt/). Applying a number of more advanced techniques to this basic HMM/GMM system can result in much lower word error rates of around 26% (Hain et al. 2007). The advances basically fall in three main areas: speaker adaptation, discriminative training, and long-context acoustic features.

Speaker adaptation: Vocal tract length normalisation (VTLN) and maximum likelihood linear regression (MLLR) have both been productively applied to adapt the basic system to a particular

talker. In the case of VTLN, a single parameter, the warp factor, is estimated (by maximum likelihood) to warp the frequency spectrum, to take account of the effect of varying vocal tract length (Cohen et al. 1995, Hain et al. 1999). It can be shown that this feature transform can (under certain assumptions) be related to a linear transform of the model parameters. A more general technique based on this idea is MLLR, in which the model parameters are linearly transformed, with the transform again estimated by maximum likelihood (Leggetter and Woodland 1995, Digalakis et al. 1995). In both cases, the efficiency of the adaptation can be improved by carrying adaptation at training time, so that training is carried out on speaker normalised models, a technique referred to (in the case of MLLR) as speaker adaptive training.

Discriminative training: The parameters of a basic HMM/GMM acoustic model are estimated using maximum likelihood estimation. Nadas (1983) and Bahl et al. (1986) noted that a directly discriminative optimality criterion can lead to lower error rates when the ideal conditions of model correctness and infinite training data do not hold. Following work by Woodland and Povey (2002) and Povey and Woodland (2002), discriminative techniques have been successfully applied to large scale speech recognition tasks. In particular they developed the minimum phone error (MPE) approach, which is a minimum Bayes risk approach that incorporates into its cost function the string edit distance between the recognised and reference strings.

Long-context features: HMM/GMM systems typically use estimates of the first and second derivatives of the acoustic features, typically estimated using a seven frame window. These features integrate local temporal context in order to provide an estimate of the dynamics of the acoustic features. More recent approaches which attempt to add context to features include feature-based MPE (Chen et al. 2006), based on a projection from the high-dimensional space of Gaussian probability densities, and the left-context, right-context (LCRC) features of Grezl et al. (2007). The LCRC features are based on two neural networks, each of which estimate the phone posterior probability of the current frame based on 25 frames left context and 25 frames of right context respectively, followed by a third "merger" neural network.

The effect of discriminative training and long context features on the word error rate (in the RT07 evaluation) is graphed in figure 3, where maximum likelihood, MPE and fMPE are employed with and without the LCRC long-context features. It can be seen that discriminative training results in a significant error reduction, with further improvements coming from the employment of both fMPE and the LCRC features.

In addition to these main techniques a number of other techniques have also been used to lower error rates including novel acoustic parameterisations based on pitch adaptive features (Garau and Renals 2008), estimation of n-gram language models well-matched to the target domain by augmenting training data using documents obtained from the web by searching with n-grams obtained from meeting transcripts (Wan and Hain 2006, Bulyko et al. 2007), and acoustic segmentation approaches specifically optimised for meetings (Wrigley et al. 2005, Dines et al. 2006). Acoustic segmentation and speech/non-speech detection remains an important problem, with nearly 10% of errors in our current system resulting from errors in the speech/non-speech detection component.

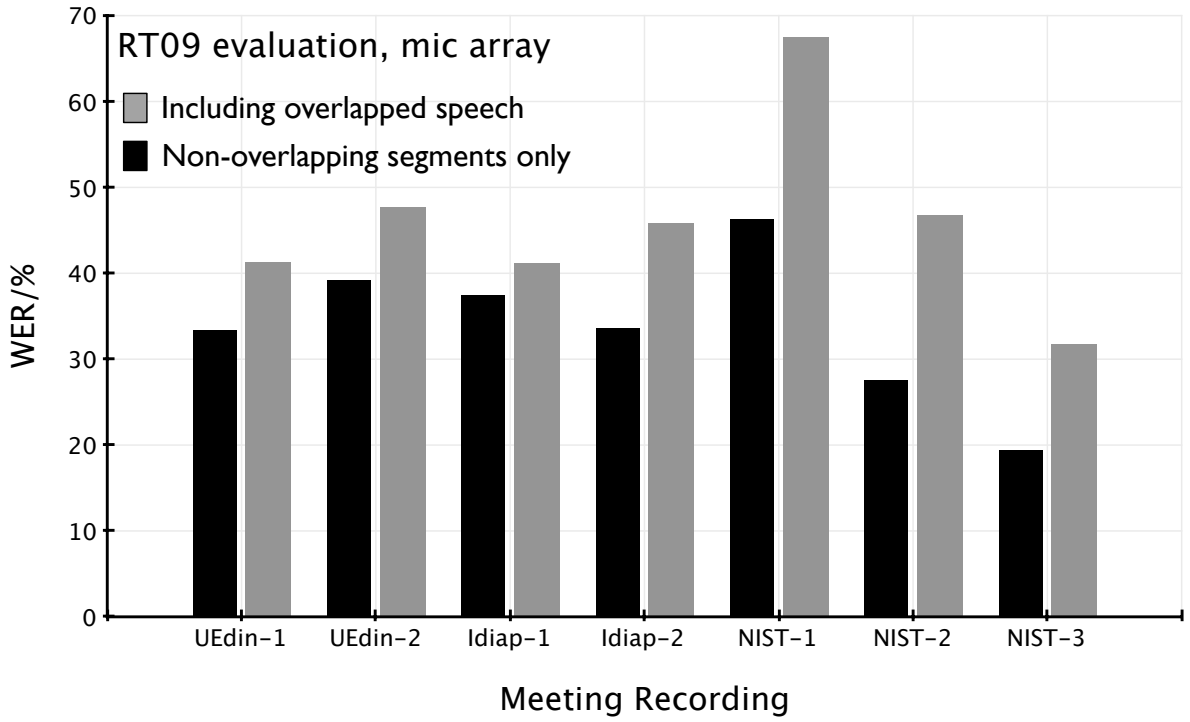

Fig. 4. Word error rates by the AMI-ASR system on the RT09 evaluation data from meeting recordings using multiple distant microphones (microphone arrays). The graph shows results from 7 meetings recorded in three different meeting rooms (UEdin, Idiap, NIST). Word error rates are compared for overlapping and non-overlapping segments.

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 mel frequency cepstral coefficients and LCRC features) are cross-adapted, and different feature representations are also combined using linear transforms such as heteroscedastic linear discriminant analysis (HLDA) (Kumar and Andreou 1998).

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 realtime system (with an error rate of about 41% for microphone array input) (Garner et al. 2009), based on an open source runtime system, Juicer (http://juicer.amiproject.org/).

The word error rates for meeting speech recognition in the NIST RT09 evaluation are shown in figure 4 for a microphone array system for the seven different meetings that formed the test set. It is clear that there is a high degree of variability across the test sets; for example, the NIST-1 meeting is a planning meeting for a social event, rather different in style to the research and design meetings that form the rest of the test set. Secondly, the effect of overlapped speech on current systems is clearly shown in this graph; if all overlapping segments are considered (including every backchannel

from every participant), than the overall word error rate is much higher compared with when only non-overlapped segments are considered. Unsurprisingly, many of the increased errors in the former case arise from deletion errors, when the overlapped segments are not detected.

4 MEETING CONTENT EXTRACTION

Richer knowledge of meeting content is required for many applications, which require meetings to be structured and distilled, and key events to be extracted. In this section we discuss the automatic segmentation of meetings, meeting summarisation and the extraction of social information such as agreement.

4.1 Segmentation

Many different ways to temporally segment a meeting have been explored—by talker, by "meeting phase", by dialogue act, or by topic—and both supervised and unsupervised methods have been used. A major commonality that has been found across all these levels of segmentation is the need for multimodal features: text (speech transcriptions), prosody (timing and intonation), various features characterising interaction patterns, and various video features. As a specific example we describe a study in which "meeting actions" are automatically detected using from multiple streams of multimodal features (Dielmann and Renals 2007).

In this study meeting actions correspond to phases of the meeting in terms of the predominant group activity. In the case of the four-person meetings that we studied in this work eight meeting actions were defined: monologues (per participant), discussion, presentation, speaking at whiteboard, and notetaking. Information about these actions is factored across the meeting participants, the outputs of sensors, and the outputs of different multimodal recognisers. Four feature streams were used: Prosody (F0, rate of speech, energy); Speaker turn features (speech activity in each of 6 locations, over 3 time periods); Lexical features (trigram language models for different meeting phases); and Visual features (motion intensity and direction of skin-like blobs).

The baseline system we used to model the meeting actions used an HMM/GMM for each action class, in which the hidden variable (HMM state) generates the entire set of features (figure 5, left). If all four feature streams are used, then the feature vector is a concatenation of the four streams, an approach referred to as early integration in the data fusion literature. The quality of the system was measured using action error rate, which is defined analogously to word error rate and based on the string edit distance between the recognised sequence of actions and the reference sequence of actions. Using a single feature stream resulted in action error rates ranging from 49–60%; combining all the feature streams into a single feature vector reduced the action error rate to 44%.

The early integration approach thus results in a high action error rate. An alternative approach was explored in which meeting actions decomposed as sequences of hidden subactions, with each subaction being responsible for a feature stream. This was represented as a dynamic Bayesian network (DBN), as illustrated on the right of figure 5. The DBN provides a much richer hidden structure, with a distributed state representation that allows feature streams to be processed independently and

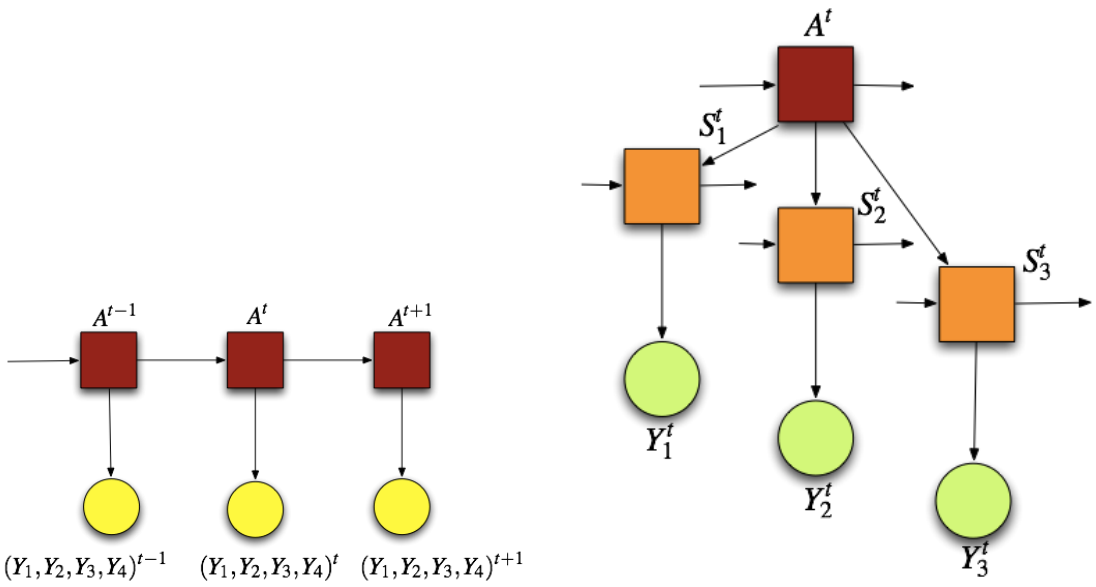

Fig. 5. Graphical models for meeting segmentation. The HMM (left) combines the multimodal features into a single feature vector; the DBN (right) processes the multimodal feature streams asynchronously and independently using a hierarchical state space in which the action state sequence is decomposed into subaction state sequences.

asynchronously. Results on same task using a three-stream DBN, with five subactions per stream, resulted in a significantly lowered action error rate of 13%. Al-Hames et al. (2006) compares a number of feature representations and models for this task.

4.2 Summarisation

Summarisation can be an extremely useful way to process a meeting. A meeting summary can shield a user from having to view a transcript of spontaneous meeting speech (which is hard to read, even without considering speech recognition errors), can enable the unstructured contents of a meeting to be included in an organisation's knowledge repository, and can be very beneficial in assisting in applications which require the review of meetings. An example of the latter is "decision audit", in which a sequence of meetings is reviewed in an attempt to understand how and why a certain decision was arrived at.

We have investigated a range of approaches to the extractive summarisation of meetings, in which a summary is constructed by selecting the most salient fragments across the meeting. A fragment may be a speaker turn, a dialogue act or a "speech spurt" (a sequence of speech from a single speaker bounded by silence, but not not necessarily corresponding to a linguistic unit). The algorithms that we have used for extractive summarisation rely strongly on speech transcripts, and employ methods such as $tf \cdot idf$ used in text retrieval (Zechner 2002). However, text-only methods for meeting summarisation can be significantly improved by considering information related to speaker turns and to prosody (Murray et al. 2006).

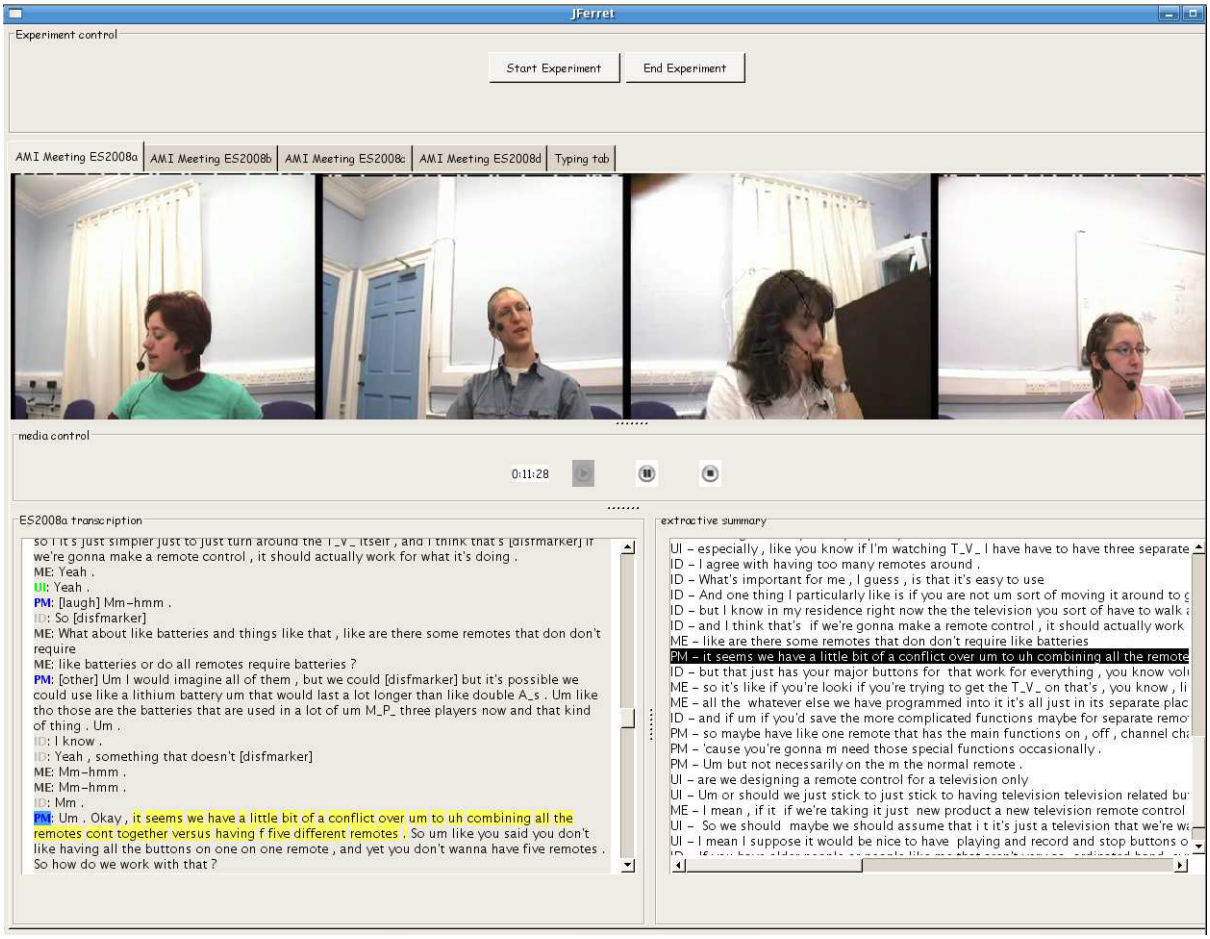

Fig. 6. Summary-based meeting browser used in decision audit evaluation of summarisation.

The evaluation of summarisation is not straightforward, since there is no single, gold-standard reference with which to compare an automatically generated summary. ROUGE (Lin and Hovy 2003) provides a way to evaluate the quality of a summary compared with multiple reference summaries, and has become widely used in text summarisation, agreeing well with subjective evaluations. However, ROUGE does not correlate well with human judgements for meeting summarisation (Liu and Liu 2008). In order to better evaluate our meeting summarisation systems, we carried out a subjective evaluation based on a decision audit task (Murray et al. 2009). Users were asked to find the factors that led to a particular decision that resulted from a sequence of four meetings in the AMI corpus. To carry out this task they were provided with meeting browsers for the four meetings, which contained transcripts, summaries and media players (figure 6). The browsers were compared based on transcript type (automatic or hand-generated) and summarisation algorithm (keywords, extractive and hand-generated). Fifty subjects participated in the evaluation and the measures were both objective (number of factors found, time taken, browser interaction statistics), and subjective based on questionnaires. We discovered that the decision audit task in itself was a challenging one for most users, in all conditions. The main results indicated that automatic summaries out-performed keyword spotting baselines, and users were able to perform the task almost as well using speech recognition transcripts, with an error

rate of about 30%, compared with using human transcripts. However there was much lower user satisfaction when using speech recognition transcripts, and users played media files significantly more, compared to when using human transcripts.

4.3 Social processing

Meetings cannot be understood purely in terms of their lexical content, since they have a clear social function. In many cases, 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) and subjective content (Wilson and Raaijmakers 2008).

Germesin and Wilson (2009) investigated the automatic detection of agreements in multi-party conversations. Again a variety of multimodal features, representing lexical, prosodic, and structural information, were employed. Using approaches based on decision trees and conditional random fields, and annotation of the AMI corpus for agreement and disagreement, it was found that agreements could be detected with up to 59% precision and 42% recall. To further emphasise the social aspect, it was also possible to automatically identify the speaker being agreed with with an accuracy of over 80% (in a four-person meeting).

5 Discussion

The meeting recognition and content extraction technologies discussed above have been integrated in a number of exemplar applications. Meeting browsers enable a meeting to be browsed based on (automatic) annotations and media files, synchronised to a common timeline. The outputs of different recognisers and sensors can be plugged into to such browsers, according to need. A summarisation-based meeting browser is shown in figure 6.

Building on the availability of online meeting speech recognition with a few seconds latency (Garner et al. 2009), we have developed some close to real-time applications, based on an architecture called "The Hub"—a real-time client/server software framework that enables annotations to be exchanged between annotation producers and consumers. Applications may be both producers and consumers, for instance most content extraction producers will also consume speech recognition output. One such application is the AMIDA Content Linking Device (Popescu-Belis et al. 2008), which performs online search and retrieval based queries generated to represent the current meeting context. In practice queries are based on the output of the speech recogniser over the previous 30s, combined with pre-specified information relating to the domain or task. The application can then use these queries to retrieve relevant documents from the web or from a repository directly related to the ongoing meeting (including previous meetings). Figure 7 shows a basic interface to content linking.

The AMI corpus was recorded using a relatively specialised set of hardware, requiring a specific room to be instrumented. More recently, the required hardware for meeting capture has become generally available, and portable meeting capture—based on a laptop, USB microphone array, and

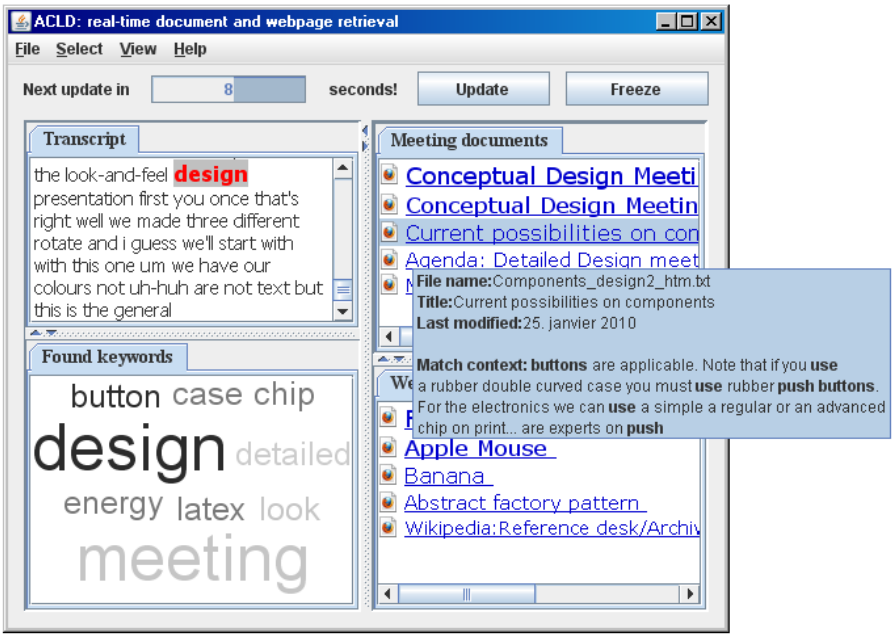

Fig. 7. Interface to the AMIDA Content Linking Device. The transcript is at the top left, a tag cloud of keywords is at the bottom right, and the left hand panels show relevant documents on the web and locally available. Mouse-over brings up a document summary.

omnidirectional camera—is now possible. Exploiting these opportunities, we have developed a prototype application for personal productivity uses, referred to as the Ambient Spotlight (Kilgour et al. 2010). In the Ambient Spotlight (figure 8), the ideas of content linking are used to integrate captured meetings with standard calendar, mail and desktop search applications.

This paper has given an overview of research into the automatic recognition of, and content extraction from, multiparty meetings. This work presents significant interdisciplinary challenges, from signal processing to discourse modelling. The AMI meeting corpus is at the centre of our attempts to address these problems, and offers an integrated, multimodal data collection with a growing number of annotations.

Acknowledgments

This work benefits from long-term collaborations with a number of people: Hervé Bourlard, Jean Carletta, Thomas Hain, Jonathan Kilgour, Mike Lincoln, Theresa Wilson, and Andrei Popescu-Belis, as well as significant contributions from a number of my current and former PhD students: Alfred Dielmann, Giulia Garau, Songfang Huang, Gabriel Murray, Le Zhang, and Erich Zwyssig. 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.

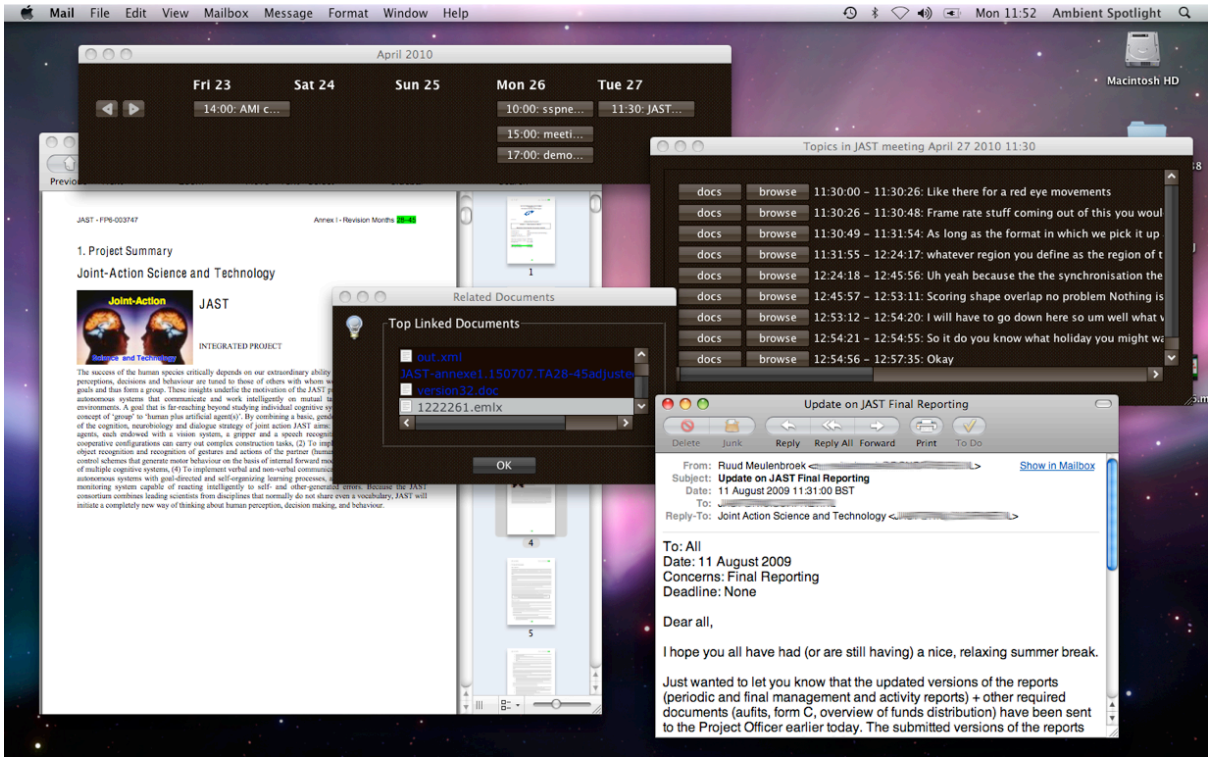

Fig. 8. The Ambient Spotlight interface. The interface is centred on a calendar display. Recorded meetings in the calendar are segmented into topics. For each topic relevant documents are retrieved.

REFERENCES

Al-Hames, M., Dielmann, A., Gatica-Perez, D., Reiter, S., Renals, S., Rigoll, G., and Zhang, D. (2006). Multimodal integration for meeting group action segmentation and recognition. In Renals, S. and Bengio, S., editors, *Proc. MLMI '05*, volume 3869 of *LNCS*, pages 52–63. Springer-Verlag.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The HCRC map task corpus. *Language and Speech*, 34(4):351–366.

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

Bahl, L., Brown, P., de Souza, P., and Mercer, R. (1986). Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *Proc IEEE ICASSP*.

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

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

Carletta, J., Evert, S., Heid, U., and Kilgour, J. (2005). The nite xml toolkit: Data model and query language. *Language Resources and Evaluation*, 39(4):313–334.

Carletta, J., Evert, S., Heid, U., Kilgour, J., Robertson, J., and Voormann, H. (2003). The NITE XML Toolkit: flexible annotation for multimodal language data. *Behavior Research Methods, Instruments,*

& Computers, 35(3):353–363.

Chen, S. F., Kingsbury, B., Mangu, L., Povey, D., Saon, G., Soltau, H., and Zweig, G. (2006). Advances in speech transcription at ibm under the darpa ears program. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1596–1608.

Cohen, J., Kamm, T., and Andreou, A. (1995). Vocal tract normalization in speech recognition: compensating for systematic speaker variability. *Journal of the Acoustic Society of America*, 97(5, Pt. 2):3246–3247.

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

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

Digalakis, V. V., Rtischev, D., and Neumeyer, L. G. (1995). Speaker adaptation using constrained estimation of gaussian mixtures. *IEEE Transactions on Speech and Audio Processing*, 3(5):357–366.

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

Gales, M. J. F., Kim, D. Y., Woodland, P. C., Chan, H. Y., Mrva, D., Sinha, R., and Tranter, S. E. (2006). Progress in the cu-htk broadcast news transcription system. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1513–1525.

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

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

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

Gatica-Perez, D., Lathoud, G., Odobez, J. M., and McCowan, I. (2007). Audio-visual probabilistic tracking of multiple speakers in meetings. *IEEE Transactions on Audio, Speech and Language Processing*, 15(2):601–616.

Germesin, S. and Wilson, T. (2009). Agreement detection in multiparty conversation. In *Proc ICMI-MLMI*.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. IEEE ICASSP*.

Grezl, F., Karafiat, M., Kontar, S., and Cernocky, J. (2007). Probabilistic and bottle-neck features for LVCSR of meetings. In *Proc IEEE ICASSP*.

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

Hain, T., Dines, J., Garau, G., Karafiat, M., Moore, D., Wan, V., Ordelman, R., and Renals, S. (2005). Transcription of conference room meetings: an investigation. In *Proc. Interspeech '05*.

Hain, T., Woodland, P. C., Niesler, T. R., and Whittaker, E. W. D. (1999). In *Proc IEEE ICASSP*.

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

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). The ICSI meeting corpus.

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

Kilgour, J., Carletta, J., and Renals, S. (2010). The Ambient Spotlight: Queryless desktop search from meeting speech. In *Proc ACM Multimedia 2010 Workshop SSCS 2010*.

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

Lee, D., Erol, B., and Graham, J. (2002). Portable meeting recorder. *ACM Multimedia*.

Leggetter, C. J. and Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, 9(2):171–185.

Lin, C. and Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. NAACL/HLT*.

Liu, F. and Liu, Y. (2008). Correlation between rouge and human evaluation of extractive meeting summaries. In *Proc. ACL*.

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

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

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

Murray, G., Renals, S., Moore, J., and Carletta, J. (2006). Incorporating speaker and discourse features into speech summarization. In *Proc NAACL*, pages 367–374.

Nadas, A. (1983). A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 31(4):814–817.

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

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

Poppe, R. and Poel, M. (2008). Discriminative human action recognition using pairwise CSP classifiers. In *IEEE FGR*.

Povey, D. and Woodland, P. C. (2002). Minimum phone error and i-smoothing for improved discriminative training. In *Proc IEEE ICASSP*.

Renals, S. and Hain, T. (2010). Speech recognition. In Clark, A., Fox, C., and Lappin, S., editors,

Handbook of Computational Linguistics and Natural Language Processing. Wiley Blackwell.

Roy, D. M. and Luz, S. (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.

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

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

Waibel, A., Schultz, T., Bett, M., Denecke, M., Malkin, R., Rogina, I., Stiefelhagen, R., and Yang, J. (2003). SMaRT: the smart meeting room task at ISL. In *Proc IEEE ICASSP*.

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

Wilson, T. and Raaijmakers, S. (2008). Comparing word, character, and phoneme n-grams for subjective utterance recognition. In *Proc. Interspeech*.

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

Woodland, P. C. and Povey, D. (2002). Large scale discriminative training of hidden Markov models for speech recognition. *Computer Speech & Language*, 16(1):25–47.

Wooters, C. and Huijbregts, M. (2008). The icsi rt07s speaker diarization system. In *Multimodal Technologies for Perception of Humans*, number 4625 in LNCS, pages 509–519. Springer.

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

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

Zechner, K. (2002). Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.