The 2005 AMI System for the Transcription of Speech in Meetings

Thomas Hain[1], Lukas Burget[2], John Dines[3], Giulia Garau[4], Martin Karafiat[2], Mike Lincoln[4], Iain McCowan[3], Darren Moore[3], Vincent Wan[1], Roeland Ordelman[5], and Steve Renals[4]

[1] Department of Computer Science,
University of Sheffield, Sheffield S1 4DP, UK.
[2] Faculty of Information Engineering,
Brno University of Technology,Brno, 612 66, Czech Republic .
[3] IDIAP Research Institute, CH-1920 Martigny, Switzerland.
[4] Centre for Speech Technology Research,
University of Edinburgh, Edinburgh EH8 9LW, UK.
[5] Department of Electrical Engineering
University of Twente, 7500AE Enschede, The Netherlands.

Abstract. In this paper we describe the 2005 AMI system for the transcription of speech in meetings used in the 2005 NIST RT evaluations. The system was designed for participation in the speech to text part of the evaluations, in particular for transcription of speech recorded with multiple distant microphones and independent headset microphones. System performance was tested on both conference room and lecture style meetings. Although input sources are processed using different front-ends, the recognition process is based on a unified system architecture. The system operates in multiple passes and makes use of state of the art technologies such as discriminative training, vocal tract length normalisation, heteroscedastic linear discriminant analysis, speaker adaptation with maximum likelihood linear regression and minimum word error rate decoding. In this paper we describe the system performance on the official development and test sets for the NIST RT05s evaluations. The system was jointly developed in less than 10 months by a multi-site team and was shown to achieve competitive performance.

1 Introduction

Transcription of speech recorded in meetings has been the focus of attention for speech researchers for quite some time. However the complexity of the input puts considerable strain on the performance of such systems. Besides the acoustic complexity, the variety of input sources and the moving speaker problems, the transcription of spontaneous speech itself is complex and normally yields results above 15% word error rate (WER). Speech transcripts of meetings are not only of interest in their own right, but are an important input for higher-level processing. Projects like AMI (Augmented Multiparty Interaction) aim to investigate the use of machine based techniques to aid people in and outside of meetings to

efficiently access meeting content. Meetings are an audio visual experience by nature, information is presented for example in the form of presentation slides, drawings on boards, and of course by verbal communication. The automatic transcription of speech in meetings is of crucial importance for meeting analysis, content analysis, summarisation, and analysis of dialogue structure.

As is often the case work on automatic recognition of speech in meetings is stimulated by yearly performance evaluations by the U.S. National Institute of Standards and Technology (NIST) [18]. Large scale work on conference room type meeting speech was initially facilitated by the collection of the ICSI meeting corpus [12] which was followed by trial NIST meeting transcription evaluations in Spring 2002. Further meeting resources were made available by NIST [8], Interactive System Labs (ISL) [2] and the Linguistic Data Consortium for the RT04s Meeting evaluations [18].

In this paper we describe the 2005 AMI system for the transcription of speech in meetings used for participation in the 2005 NIST RT evaluations (RT05s). The system was designed for participation in the speech-to-text part of the evaluations, in particular transcription of speech recorded with multiple distant microphones (MDM), the primary test condition, and individual headset microphones (IHM). Both input sources are processed using different front-ends, however the recognition process is based on a unified system architecture. The RT05s evaluations differ from those of previous years in that tests are conducted both on meetings in conference room style and lecture room style. The system presented here has been developed solely for the purpose of transcribing conference room style meetings, with the same system being used for the transcription of the lecture room meeting data[6]. Data from new sources have further enhanced the richness of the testing conditions in terms of input speech, recording conditions and content. The new data originates from data collection efforts as part of two European projects, AMI[7] and CHIL (Computers in the Human Interaction Loop[8]) as well as from collections at the Virginia Polytechnic and State University.

The rest of the paper is structured as follows: First we describe the data resources used followed by a description of our generic system architecture and the main system components, including an analysis of the performance of various components on the RT05s evaluation data sets. In following sections we give an overview of the complete system and its passes. This is contrasted with results using manual segmentation.

2 Meeting Resources

The ICSI Meeting corpus [12] is the largest meeting resource available consisting of 70 technical meetings at ICSI with a total of 73 hours of speech. The number of participants is variable and data is recorded with head-mounted and a

[6] This excludes the use customised language models, see Section 4.4. For that reason we do not specifically report results on lecture room data unless required.

[7] See `http://www.amiproject.org`

[8] See `http://chil.server.de`.

total of four table-top microphones. We have not used any other microphones present in the room. Further meeting corpora were collected by NIST [8] and ISL [2], with 13 and 10 hours respectively. Both NIST and ISL meetings have unconstrained content (e.g. people playing games or discussing sales issues) and variable number of participants. In our development we made use of the official RT04s development and evaluation sets (rt04sdev and rt04seval). Both sets include 10 minute extracts from 8 meetings recorded at the 3 sites above and the Linguistic Data Consortium (LDC). As part of the AMI project a major collection and annotation effort of the AMI meeting corpus[3] is currently underway. Data is collected at three different instrumented meeting rooms in Europe (Edinburgh, IDIAP, TNO). The target size of the corpus is more than 100 hours of transcribed speech. The meeting language is English, but many participants are non-native speakers of the language. Each meeting normally has four participants and the corpus will be split into a scenario portion and an unconstrained meetings portion. Each scenario in the corpus consists of four meetings with the same participants working on a constrained task. For the benefit of the RT05s evaluations, AMI has released a preliminary development set (rt05samidev) and approximately 16 hours of scenario training data. In this work both resources were used.

For the purpose of development of systems for transcription of lecture room speech a development set (rt05slectdev) was provided by CHIL. However this was provided very late and due to time constraints could only be used for language model (LM) optimisation. In this paper we further report results on the RT05s evaluation sets from the conference room and lecture room data (rt05seval and rt05slecteval respectively). Both sets are based on 10 minute extracts from individual meetings. The IHM and MDM tests are conducted on the same 10 minute extract.

3 System Architecture

The system architecture overview presented in this section is generic to both the IHM and MDM systems. A more detailed description of system components is provided in the following section. The IHM and MDM systems differ only in the processing of the input audio and the use of input source specific acoustic models in the various processing stages. The system operates in a total of 6 passes. Figure 1 shows a schematic representation of the processes. In the first pass (P1) the input data is segmented and transformed into a stream of 39 dimensional MF-PLP feature vectors[22]. Speech segments have a start and an end time as well as a channel/speaker label. A first recognition pass is conducted with acoustic models trained using maximum likelihood estimation (MLE) and a trigram LM (see Section 4.4). The resegmented output of this pass is used only for estimation of the vocal tract length normalisation (VTLN) warp factors on a per input channel basis. In the second pass (P2) the VTLN warp factors are determined and the audio data is recoded with these warp factors. Then a second decoding pass with acoustic models trained on VTLN data is performed. The P2 acoustic modelling includes a smoothed heteroscedastic linear discriminant analysis

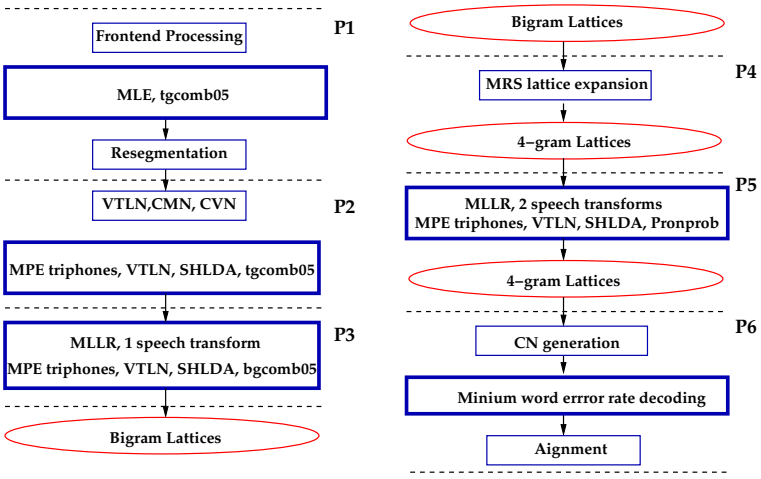

Fig. 1. Processing stages of the 2005 AMI meeting transcription system.

(SHLDA) input transform[15] and acoustic models are trained (in the IHM case) using the minimum phone error(MPE) criterion[20]. The output of P2 is used to adapt the acoustic model means and variances using maximum likelihood linear regression [7]. Two transforms, one for speech and one for silence are estimated. A third decoding pass (P3) uses MLLR adapted P2 models to generate bigram lattices. As all subsequent stages only process lattices to constrain the search space the use of a bigram in P3 avoids too harsh constraints.

In pass P4, the bigram lattices are first expanded using a trigram language model, followed by a second expansion using 4-gram LMs. For conference room data this expansion uses language models optimised for each meeting resource (MRS). The 4-gram lattices generated in P4 are used for rescoring in the following pass P5. Here models are adapted using up to two speech transforms using a regression class tree. Lattice rescoring further makes use of pronunciation probabilities estimated on the training data [11]. The output of this pass is a set of lattices which form the input to the final pass, P6. Here confusion networks [16] are formed and the most probable word from each confusion set is selected. The final output is then aligned using the P5 acoustic models.

4 System Components

In this section a more detailed discussion of the system components as outlined in Section 3 is presented. First a brief description of the front-end blocks, both for the IHM and MDM cases is given. This is followed by a description of acoustic and language model training.

4.1 Front-end Processing

A common system architecture was chosen for both IHM and MDM sub-systems. This was possible due to the enhancement based setup chosen for MDM processing. In both cases the descriptions below do not include the feature extraction process. For more details the reader is referred to [10].

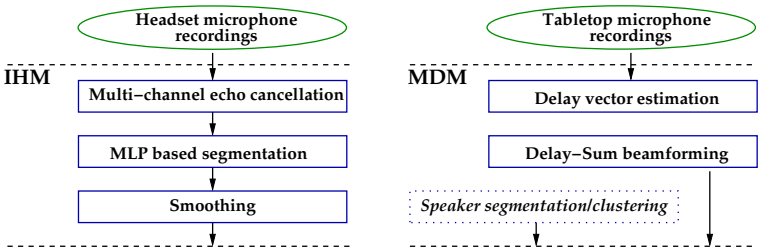

Fig. 2. Front-end processing of IHM and MDM data.

Individual Headset Microphone Processing The main task for the front-end processing of IHM data is speech activity detection (SAD). Figure 2 outlines the processes involved. First cross talk suppression is performed at the signal level using adaptive-LMS echo cancellation[17]. Additions to the basic system are: the use of multiple reference channels in cancellation; automatic estimation and correction of skew between channels; automatic cross-talk level estimation; and ignoring of channels which produce low levels of cross-talk. Updates are further made on a per sample basis to account for non-stationary 'echo' path.

The SAD system used here is a straight-forward statistical based approach with additional components to control cross-talk between channels. A 14 dimensional MF-PLP [22] feature vector is augmented with additional features: normalised RMS energy, signal and spectrum kurtosis, and as a voicing strength measure based on the maximum amplitude in the speech cepstrum in the range of frequencies 50-300Hz [19, 23]. A Multi-Layer-Perceptron (MLP) with a 31 frame input layer, a 5 unit hidden layer and an output layer of two classes is trained. Ten meetings from each meeting resource serve as training data totalling to around 20 hrs of data. A further five meetings from each corpus are used to determine early stopping of the parameter learning. The utterance segmentation uses Viterbi decoding with scaled likelihoods and a minimum segment duration of 0.5 seconds. In a final processing step the output of the segmenter is smoothed by padding segments with 0.1 seconds, merging overlapping segments in the process. Table 1 shows frame error rate results on the rt05seval before and after segmentation. Note that the relationship between false alarm and false reject rates differs substantially between meeting resources. The performance overall on the test data shows relatively high false reject rates. Smoothing the segment boundary estimates by padding allows to reduce the false reject rates significantly.

Multiple Distant Microphone Processing The basic processing stages of MDM processing are outlined in Figure 2. Since the position of microphones in the meeting room is not fixed for this task an approach that does not require geometry information was used.

First gain calibration is performed by normalising the maximum amplitude level of each of the input files. Then a noise estimation and removal procedure is run. This in itself is a two pass process. On the first pass the noise spectrum $\Phi_{nn}(f)$ of each input channel is estimated as the noise power spectrum of the

Table 1. Segmentation performance (in %) on rt05seval. FA denotes false acceptance, FR false reject, and speech the percentage of speech in the reference. TOT gives the overall performance whereas TOT(REL) are relative to the associated class.

	AMI	ISL	ICSI	NIST	VT	TOT	TOT(REL)
RAW							
FA	1.29	1.52	0.71	1.49	3.70	1.64	2.00
FR	4.49	3.03	3.36	2.81	1.12	2.94	16.23
speech	24.40	28.84	13.79	15.56	14.83	18.12	
SMOOTHED							
FA	1.90	2.55	1.21	2.05	4.34	2.22	2.71
FR	3.80	2.01	2.71	2.18	0.83	2.30	12.69
speech	24.40	28.84	13.79	15.56	14.83	18.12	

M lowest energy frames in the file ($M = 20$ was used. On the second pass a Wiener filter with transfer function $\frac{\Phi_{xx}(f)-\Phi_{nn}(f)}{\Phi_{xx}(f)}$ (where $\phi_{xx}(f)$ is the input signal spectrum) is applied to each channel to remove stationary noise. The noise coherence matrix Q, estimated over the M lowest energy frames, is computed. Finally delay vectors between each channel pair are calculated for every frame in the input sample. The delay between two channels is the time difference between the arrival of the dominant sound source and is calculated by finding the peak in the generalised cross correlation[13] between input frames across two channels.

The delay vector is given as the delays for all pairs with respect to a single reference channel - there are therefore N delays in each vector, with the delay for the reference channel equal to 0. Further a vector of relative scaling factors is calculated, corresponding to the ratio of frame energies between each channel and the reference channel. The start and end times in seconds, along with the delay and scaling factors are output for each frame. The delay and scaling vectors are then used to calculate beamforming filters for each frame using the standard superdirective technique [4, 5]. Segments and speaker labels were provided by SRI/ICSI[21].

While this approach is robust to a variety of configurations, for a small number of sparsely located microphones (as for some rooms in the rt05seval set) delay estimation can be unreliable and significant spatial aliasing occurs.

4.2 Acoustic Models

Acoustic models are phonetic decision tree state clustered triphone models with standard left-to-right 3-state topology. Models are trained up to 16 mixture components using MLE with standard HTK[9] procedures and contain approximately 4000 states. For more details on the training process the reader is referred to [10]. In previous experiments [10] we found that maximum a posteriori (MAP)[9] adaptation from conversational telephone speech (CTS) models gave better performance than training solely on meeting data.

VTLN was applied both in training and testing, both on IHM and MDM. For training an iterative procedure was used alternating the estimation of warp-

[9] The Hidden Markov Model Toolkit (HTK). `http://htk.eng.cam.ac.uk`.

Table 2. %WER on rt05seval IHM rescoring 4-gram lattices with pronunciation probabilities and various models. By default models are trained on meeting data only.

	TOT	Sub	Del	Ins	AMI	ISL	ICSI	NIST	VT
CTS adapted	39.1	20.0	13.4	5.7	39.9	35.1	36.0	46.9	37.6
CTS adapted, VTLN	36.9	18.5	13.0	5.5	37.0	33.1	34.4	45.2	34.8
VTLN	37.2	18.8	13.2	5.2	36.4	33.0	36.1	45.5	35.0
HLDA	35.7	17.8	13.4	4.6	36.0	31.0	33.9	43.3	34.6
SHLDA	35.6	17.7	13.3	4.5	35.6	30.3	34.5	42.8	34.7
SHLDA-MPE	32.9	15.8	13.3	3.8	32.8	27.8	32.3	39.8	31.9

ing factors and model parameter updates. For IHM initial warp factor estimates were obtained from CTS-adapted models. Experimental evidence shows improved WER performance with warp factor estimation at a reduced bandwidth of 3800Hz. Initial experiments using IHM models for warp factor estimation on MDM data yielded a performance degradation. Hence IHM VTLN models were adapted to the MDM VTLN data where a single training iteration was found to yield good results that could not be improved further.

Feature space transformation was applied in the form of smoothed heteroscedastic linear discriminant analysis (SHLDA) [15]. The transform was used to reduce a 52 dimensional feature vector (standard plus third derivatives) to 39 dimensions. HLDA estimation procedure[14] requires the estimation of full covariance matrices per Gaussian. SHLDA in addition uses smoothing of the covariance estimates by interpolating with standard LDA type within-class covariances. The adaptation of CTS models when using SHLDA is non-trival due to the reduced bandwidth of CTS data. To avoid further issues with discriminative training no CTS data was used in conjunction with SHLDA.

All further models were trained using the minimum phone error criterion [20]. The implementation of MPE used here is similar to that described in [20]. For this purpose numerator and denominator lattices were generated using the SHLDA models and a bigram LM interpolated with a unigram model that includes training set specific words. The phone times as obtained in recognition are used to improve speed in training. Only means and variances are modified and parameter update makes use of I-smoothing. Performance was found to stabilise after 10 training iterations[10].

Table 2 shows lattice rescoring results on rt05seval IHM for models of increasing complexity. Note the 0.3% performance degradation from the use of unadapted models which is compensated by 1.6% improvement from SHLDA. Another 2.8% absolute are gained by the use of MPE training. It can be observed that model improvement has little impact on the deletion rate.

4.3 Training Data Selection

Training data for IHM is given by the reference transcripts. In total 104 hours of speech were available from resources outlined in Section 2, albeit a significant

[10] Both SHLDA and MPE are developed as part of the STK HMM toolkit: `http://www.fit.vutbr.cz/speech/sw/stk.html`.

Table 3. MDM Data selection. IHM denote IHM segments (inc. overlapped speech). sil-bound and word-bound denote methods for removing overlap (cut at silence or word boundaries), sn denotes silence normalisation.ASL denotes the average segment length.

	#Segments	Size (hours)	ASL (sec)	%Silence
IHM	136822	104.27	2.74	27.0
sil-bound	84044	62.33	2.67	21.0
word-bound	94940	65.78	2.49	21.1
word-bound + sn	96086	62.96	2.36	18.0

Table 4. Size of various text corpora in million words (MW).

Corpus	#words (MW)
Swbd/CHE	3.5
Fisher	10.5
Web (Swbd)	163
Web (fisher)	484
Web (fisher topics)	156
BBC - THISL	33
HUB4-LM96	152
SDR99-Newswire	39
ICSI/ISL/NIST/AMI	1.5
Web (ICSI)	128
Web (AMI)	100
Web (CHIL)	70

proportion of the data is silence. The special processing setup for MDM data (see Section 4.1) however makes additional processing necessary as the system cannot cope with overlapped speech. A straight forward exclusion of all segments with overlaps would have resulted in removal of more than 60% of the data and hence was not an option. Table 3 compares several data selection techniques based on alignments. *sil-bound* denotes cuts at the nearest boundary where silence occurs, *word-bound* the nearest word-boundary regardless of silence. With *sn* further silence beyond 0.2 seconds at segment boundaries and within segments was removed. The word-bound+sn configuration showed marginally better performance and was used for MDM model training.

4.4 Vocabulary, Language Models and Dictionaries

The recognition vocabulary is set to cover the 50000 most frequent words using a procedure outlined in[10]. The same vocabulary was used both for lecture and conference room style meetings. Pronunciation dictionaries are based on the UNISYN pronunciation lexicon [6] which was manually augmented[10]. Pronunciation probabilities are estimated from alignment of the training data[11].

As in previous work, LMs trained on a large number of corpora were used to derive meeting room specific and generic language models by optimisation of interpolation weights. The most important corpora are listed in Table 4. A full discussion of all source material would go beyond the scope of this paper. It is important to note that a collection of data from the web using tools and methods as provided by [1] was performed using both AMI and CHIL data as the basis. In both cases the proposed approach was altered to focus on previ-

Table 5. Perplexities for 4-gram LMs on rt04dev and rt05samidev

	Language models					
Data source	ICSI	NIST	ISL	AMI	LDC	fgcomb05
ICSI	82.734	86.1662	87.3345	97.1024	109.86	84.1826
NIST	101.442	103.668	102.054	105.683	109.212	98.8722
ISL	110.124	110.99	106.66	119.327	114.483	108.588
AMI	92.9651	108.865	108.723	77.2817	101.714	84.1282
LDC	92.3824	92.761	87.6343	99.0105	84.2745	90.5354
AllDev	86.9236	93.2191	93.6604	92.0517	106.716	85.381

Table 6. %WERs on rt05seval showing the effect of CN decoding. Word times are corrected by alignment.

CN decoding	Word time correction	IHM	MDM
		32.1	44.2
	×	31.2	42.2
×		31.5	44.0
×	×	30.6	42.0

ously unobserved contexts. This approach has in particular lead to a dramatic reduction in perplexity for lecture room data by more than 30%.

Table 5 shows perplexities for language models tuned to specific meeting resources as well as in combination. It is evident the meeting room specific models outperform the combined models. Hence the lattice expansion to 4-gram lattices (see Section 3) was performed using meeting resource specific models. This gave an additional 0.5% WER reduction on the rt04seval set.

4.5 Minimum Word Error Decoding

Minimum word error rate decoding[16] is a widely used technique to counter the fact that the standard speech recognition objective function is to minimise sentence instead of word error rate which is the measurement metric. Table 6 compares the performance both on IHM and MDM. In both case the gain from this technique was found to be moderate. The table also shows the effect of correcting the word times by alignment. Standard decoding adds between-word silence to the end of a word, thus artificially lengthening words. Secondly, confusion network decoding uses heuristic rules to define word times. Hence again re-alignment is needed to correct the times.

5 Overall System Performance

Table 7 shows WER results for the 2005 AMI meeting transcription system on a per pass basis. The result for P3 is higher than that for P2 due to the use of a bigram language model. The major reduction in WER at P6 can be explained by the use of alignment (see above). The high deletion rate is a main contributor to the error rate. Overall the WER reduction up to P6 is 10.5% absolute, however most of the gain is already obtained in P2. The associated results on rt05seval MDM are shown in Table 8. Note that a similar improvement is obtained to that observed on IHM data, again with relatively high deletion rates. Particularly poor performance on VT data has a considerable impact on performance (only 2 distant microphones!).

Table 7. %WER on rt05seval IHM.

	TOT	Sub	Del	Ins	Fem	Male	AMI	ISL	ICSI	NIST	VT
P1	41.1	21.1	14.7	5.3	41.1	37.2	42.3	36.3	37.1	49.1	41.1
P2	33.1	15.9	13.4	3.9	33.1	28.2	33.4	27.2	32.8	39.5	32.8
P3	34.4	16.9	13.7	3.9	34.4	28.7	34.8	27.7	33.5	41.8	34.6
P4.tg	32.2	15.3	13.1	3.8	32.2	27.3	32.3	26.1	32.1	39.3	31.4
P4.fg	32.3	15.5	12.9	3.9	32.3	27.7	32.6	26.4	31.9	39.5	31.2
P5	32.1	15.3	12.8	4.0	32.1	27.4	32.7	26.3	31.8	39.1	30.5
P6	30.6	14.7	12.5	3.4	30.6	25.9	30.9	24.6	30.7	37.9	28.9

Table 8. %WER on rt05seval MDM.

	TOT	Sub	Del	Ins	Fem	Male	AMI	ISL	ICSI	NIST	VT
P1	53.6	32.1	17.3	4.1	53.6	56.4	46.5	50.2	48.2	53.6	63.0
P2	50.8	31.3	14.8	4.7	50.8	51.4	44.7	46.7	43.6	51.6	60.4
P3	50.4	31.1	14.6	4.7	50.4	53.0	44.7	47.0	45.2	48.9	59.7
P4.tg	48.4	30.0	13.6	4.8	48.4	49.4	43.9	44.8	42.5	46.9	57.2
P4.fg	47.9	29.5	13.7	4.7	47.9	49.3	42.4	45.0	41.8	47.4	56.6
P5	44.2	26.0	14.0	4.1	44.2	42.6	38.6	38.9	39.2	43.8	53.2
P6	42.0	25.5	13.0	3.5	42.0	42.0	35.1	37.1	38.4	41.5	51.1

5.1 Manual Segmentation

In previous sections we have shown that automatic segmentation is still a main source of error. Table 9 compares results with reference and automatic segmentation. Both on MDM and IHM the automatic segmentation naturally increases deletion rates, however the effect is far stronger on IHM where the overall difference between automatic and manual segmentation is 6.4%. The gain from confusion network decoding is further decreased with automatic segmentation. The absolute gain from P1 to P6 is similar in absolute terms, with or without manual segmentation.

5.2 Lecture Room Meetings

Lecture room meetings as included in the RT05s evaluations originate only from one recording site. Presentation sessions are mixed with question/answer meetings where more than one speaker talks. In this work no development work was performed due to lack of time. The system for conference room meetings was used as described except for language models optimised on the associated development data with additionally collected web-data. For MDM transcription only the four microphones on the table were used. Table 10 shows WERs both on IHM and MDM recordings. It is interesting to note that the WERs are in the same range as on lecture room data, however the overall gain of the passes is larger. Deletion rates are considerably lower on IHM compared to the results on conference room data.

6 Conclusions

This is the first participation of the AMI-ASR team in NIST evaluation and the system presented here was developed from scratch in less than 10 months

Table 9. %WER summary for rt05seval

	IHM				MDM			
	refseg		autoseg		refseg		autoseg	
	TOT	Del	TOT	Del	TOT	Del	TOT	Del
P1	34.9	7.1	41.1	14.7	50.6	11.8	53.6	17.3
P2	26.0	7.1	33.1	13.4	46.4	11.4	50.8	14.8
P3	27.4	7.4	34.4	13.7	47.8	12.5	50.4	14.6
P4	24.5	6.4	32.3	12.9	45.1	11.5	47.9	13.7
P5	24.5	6.3	32.1	12.8	42.0	12.2	44.2	14.0
P6	24.2	6.4	30.6	12.5	40.7	12.3	42.0	13.0

Table 10. %WER on rt05slecteval.

	IHM				MDM			
	TOT	Sub	Del	Ins	TOT	Sub	Del	Ins
P1	44.4	26.4	5.0	12.9	65.0	47.6	9.9	7.5
P2	33.0	19.1	5.2	8.7	60.0	43.4	10.0	6.7
P3	33.7	19.7	5.3	8.6	59.9	43.0	11.0	5.9
P4	31.4	18.2	4.8	8.3	58.8	42.2	10.1	6.5
P5	31.1	18.2	4.6	8.3	54.8	38.7	11.2	5.0
P6	30.4	17.7	4.6	8.0	53.5	37.2	11.6	4.7

in a joint multi-site effort. The system was shown to yield very competitive performance for the transcription of meeting data in the NIST RT05s evaluation both on lecture and conference room data. We have also described and analysed a series of potential short-comings that will be addressed in the future. Particular emphasis will be placed on improving the IHM and MDM front-end processing.

Acknowledgements

This work was largely supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811). The authors thank the rest of the AMI-ASR team for their valuable contributions: Barbara Peskin, Jan Cernocky, Jithendra Vepa, and Chuck Wooters. We also would like to thank Andreas Stolcke and ICSI for providing the segments and speaker labels for MDM data, and the Cambridge University Engineering Department for providing the h5train03 CTS training set and for the right to use Gunnar Evermann's HDecode at the University of Sheffield.

References

1. I. Bulyko, M. Ostendorf, and A. Stolcke (2003). Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures. in Proc HLT'03.
2. S. Burger, V. MacLaren, H. Yu (2002). The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style. In Proc. ICSLP'2002.
3. J. Carletta,S. Ashby, S. Bourban, M. Guillemot M. Kronenthal, G. Lathoud, M. Lincoln, I. McCowan, T. Hain, W. Kraaij, W. Post, J. Kadlec, P. Wellner, M. Flynn, and D. Reidsma (2005). The AMI Meeting Corpus. In Proc. MLMI'05, Edinburgh.

4. H. Cox, R. Zeskind, and I. Kooij (1986). Practical supergain. IEEE Trans. ASSP, Vol 34(3), pp 393–397.
5. H. Cox, R. Zeskind, and M. Owen (1987). Robust adaptive beamforming. IEEE Trans. ASSP, Vol. 35(10), pp. 1365–1376.
6. S. Fitt (2000). Documentation and user guide to UNISYN lexicon and post-lexical rules, Tech. Rep., Centre for Speech Technology Research, Edinburgh.
7. M.J.F. Gales and P.C. Woodland (1996). Mean and Variance Adaptation within the MLLR Framework. *Computer Speech & Language*, Vol. 10, pp. 249–264.
8. J.S. Garafolo, C.D. Laprun, M. Michel, V.M. Stanford, and E. Tabassi (2004). In Proc. 4th Intl. Conf. on Language Resources and Evaluation (LREC'04).
9. J.L. Gauvain and C. Lee (1994). MAP estimation for multivariate Gaussian mixture observation of Markov Chains, IEEE Tr. Speech& Audio Processing, Vol. 2, pp. 291–298.
10. T. Hain, L. Burget, J. Dines, I. McCowan, G. Garau, M. Karafiat, M. Lincoln, D. Moore, V. Wan, R. Ordelman and S. Renals (2005). The Development of the AMI System for the Transcription of Speech in Meetings, In Proc. MLMI'05, Edinburgh.
11. T. Hain (2005), Implicit modelling of pronunciation variation in automatic speech recognition. Speech Communication, Vol. 46(2), pp. 171–188.
12. A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke and C. Wooters (2003). The ICSI Meeting Corpus. In Proc. ICASSP'03, Hong Kong.
13. C. H. Knapp and G. C. Carter (1976). The generalized correlation method for estimation of time delay/ IEEE Transactions on Acoustics, Speech and Signal Processing, Trans. ASSP, Vol 24, pp 320–327.
14. N. Kumar (1997), Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition.PhD thesis, John Hopkins University, Baltimore.
15. L. Burget (2004), Combination of Speech Features Using Smoothed Heteroscedastic Linear Discriminant Analysis. in Proc. ICSLP'04, p 4–7, Jeju Island, Korea.
16. L. Mangu, E. Brill and A. Stolcke (1999). Finding Consensus Among Words: Lattice-Based Word Error Minimization. In Proc. Eurospeech'99, pp. 495–498, Budapest.
17. D. Messerschmitt, D. Hedberg, C. Cole, A. Haoui and P. Winship (1989). Digital voice echo canceller with a TMS32020. Appl. Rep. SPRA129, Texas Instruments.
18. Spring 2004 (RT04S) Rich Transcription Meeting Recognition Evaluation Plan. NIST, US. Available at `http://www.nist.gov/speech`.
19. T. Pfau and D.P. W. Ellis (2001). Hidden Markov model based speech activity detection for the ICSI meeting project. Eurospeech'01.
20. D. Povey and P.C. Woodland (2002), Minimum Phone Error and I-Smoothing for Improved Discriminative Training, In Proc. ICASSP'02, Orlando.
21. A. Stolcke, C. Wooters, N. Mirghafori, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin and M. Ostendorf (2004). Progress in Meeting Recognition: The ICSI-SRI-UW Spring 2004 Evaluation System. In Proc. NIST RT04S Workshop.
22. P.C. Woodland, M.J.F. Gales, D. Pye and S.J. Young (1997). Broadcast News Transcription using HTK. In Proc. ICASSP'97, pp. 719-722, Munich.
23. S. Wrigley, G. Brown, V. Wan and S. Renals (2005). Speech and crosstalk detection in multichannel audio. In IEEE Trans. Speech& Audio Proc., Vol 13(1), pp. 84–91.