The 2007 AMI(DA) System for Meeting Transcription

Thomas Hain[1], Lukas Burget[2], John Dines[3], Giulia Garau[4], Martin Karafiat[2], David van Leeuwen[5], Mike Lincoln[43], and Vincent Wan[1]

[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] TNO,2600 AD Delft, The Netherlands

th@dcs.shef.ac.uk

Abstract Meeting transcription is one of the main tasks for large vocabulary automatic speech recognition (ASR) and is supported by several large international projects in the area. The conversational nature, the difficult acoustics, and the necessity of high quality speech transcripts for higher level processing make ASR of meeting recordings an interesting challenge. This paper describes the development and system architecture of the 2007 AMIDA meeting transcription system, the third of such systems developed in a collaboration of six research sites. Different variants of the system participated in all speech to text transcription tasks of the 2007 NIST RT evaluations and showed very competitive performance. The best result was obtained on close-talking microphone data where a final word error rate of 24.9% was obtained.

1 Introduction

Transcription of meetings is an interesting task for a wide range of communities. Apart from face-to-face meetings, telephone conferences are currently replaced by extended video conferencing which brings new challenges to supportive technologies. Presentations and lectures are more and more available, for streamed or static consumption or even in the context of global virtual worlds. In most of these cases it is clear that the lack of personal presence is a deficiency that hinders both in observation and action. For this purpose meeting analysis, such as conducted in the AMIDA project[6] is important for all of these tasks. In particular the transcription of the spoken words in meetings is vital to allow higher level processing, such as for example summarisation or addressee detection.

[6] See http://www.amiproject.org.

NIST evaluations on meeting data have been conducted regularly since 2002. During the past two years the meeting domain was split into conference and lecture meeting parts. This year a new component, so-called coffee breaks were added. Whereas the former two are clearly relatively formal events, the idea of the latter was to incorporate more informal settings that would include more lively interactions. There are several other ways by which meeting transcription can be investigated. The most obvious one, a distinction by recording method, i.e. the microphone sources, has been made in NIST evaluations from the start. More recently different ways of measuring performance were added. Whereas until 2005 word error rates (WERs) were measured on speech sections where only a single speaker is talking, more recently speech from multiple voices at the same time is scored. Furthermore, in 2007 the Speaker Attributed Speech To Text (SASTT) task was added where a word also must carry speaker information. A correct word is counted as mis-recognised if it carries an incorrect speaker label and vice versa. Hence naturally SASTT error rates are necessarily at least as high as STT scores.

The development of the AMI(DA) system for meeting transcription is a joint development by several research institutions associated with the AMI/AMIDA projects under the leadership of the University of Sheffield. The first system for participation in NIST evaluations was built in 2005 and the group has contributed data and participated in RT evaluations since, with important help of the ICSI/SRI team on segmentation and data issues. The 2007 system is no exception apart from the fact that the AMI(DA) system now provides all components necessary for transcription of IHM and MDM data, both for conference and lecture room meetings, and STT and SASTT sub-tasks [?]. In this paper we describe the starting point of our work this year, new developments, and the final system architecture, including detailed performance results on the 2006 and 2007 RT evaluation test sets.

2 The AMI Systems

The first AMI system for meeting transcription was developed in 2005 [?,?]. The most important technological features were the use of the UNISYN dictionary [1], smoothed heteroscedastic linear discriminant analysis (S-HLDA) [?], adaptation from conversational telephone speech (CTS) data, vocal tract length normalisation (VTLN), discriminative training using the minimum phone error criterion (MPE)[?], and maximum likelihood linear regression (MLLR) adaptation[?]. The system was based on HTK[7] and the SRI language modelling toolkit[8].

In 2006 the system was improved significantly in several areas: automatic segmentation of audio; improved language model data collection; improved acoustic modelling incorporating SAT and adaptation from CTS; and posterior-based

[7] The Hidden Markov Model Toolkit. See ‘://htk.eng.cam.ac.uk.

[8] See http://www.speech.sri.com/projects/srilm.

Figure 1. The AMI 2006 RT System architecture. P1-P6 denote system passes, M1-M3 denote different acoustic models which are approximately described in the boxes, all other acronyms can be found in the main text. Re-segmentation is arefinement of the initial output of speech/non-speech detection.

feature extraction[**?**]. Figure 1 shows the associated system diagram. The system operated in several passes where the initial pass only served for adaptation purposes. Lattices are generated in the second and third passes that are later rescored with different acoustic models. Note that the system also includes confusion network (CN) generation. Originally system combination of the outputs yielded, if any, only minor improvements, and hence the final submitted system included only one branch of the rescoring network. Transcription systems for IHM and MDM were equivalent, the MDM system architecture was a subset of the IHM one. The only difference between IHM and MDM modules, apart from the front-end processing, was the training data for acoustic models. Table 1 shows the system performance of the 2006 system on the 2006 conference evaluation data set (*rt06seval*). On IHM data the final WER was 24.2% , however on MDM (on the same meetings!) the result was 40.9% (For a detailed description of IHM and MDM please refer to [**?**]). The table also shows first (P1) and third (P3) pass performance. The P1 performance is considerably poorer due to the fact that decoding is performed with unadapted maximum likelihood (ML) trained models. The third pass result is already much closer to the final result and can be obtained with about a third of the processing time (the complete system runs in 60-100 × real-time). Note also that the performance is very similar for all meeting data sources (See Section 3.3 for more detail).

3 New developments in 2007

In the following we give a short overview of the developments that lead to the 2007 AMIDA RT meeting system. Major improvements originate from an in-

Table 1. AMI RT 2006 system performance in %WER on the *rt06seval* test set for the most important passes. Results for both IHM and MDM are shown. CMU/EDI/NIST/TNO/VT denote the different meeting corpora that are part of the test set.

IHM	**TOT**	CMU	EDI	NIST	TNO	VT	**MDM**	**TOT**
P1	**42.0**	41.9	41.0	39.0	42.1	44.8	P1	**58.2**
P3	**26.0**	25.7	24.6	25.2	26.3	29.5	P3	**42.0**
P5a-CN	**24.2**	24.0	22.2	23.2	23.6	28.2	P4a-CN	**40.9**

creased amount of training data as well as several small, but significant changes such as a word list cleaning, improved language modelling procedures and data acquisition, enhanced adaptation from CTS, stream-lined discriminative training procedures, new and modified front-ends and system combination.

3.1 Word lists

The selection of word lists and dictionaries for recognition has a considerable impact on performance. Whereas the choice of a pronunciation has an obvious direct effect, the choice of what constitutes a word, i.e. a sensible unit for language models, is often neglected. The wide-spread use of hyphenation of words causes entries in dictionaries which essentially are duplicates. These can only be differentiated by a (then) fixed language model score. Other unusual orthographic forms and mispronunciations also can be found frequently in conversational speech databases and data collected from the Internet. In order to improve the word lists generated for the AMI systems the following procedure was developed: Words for which pronunciations exist (including partial words) are sorted into groups according to their quality, ranging from high quality Q5 to low quality Q1. Words are assigned scores subjectively on the following basis:

Q5 Can be found in a dictionary or encyclopedia
Q4 A spell check will accept these words given less strict settings.
Q3 Variations of words that are unusual, perhaps not in a dictionary or incorrectly conjugated, but may still exist in speech. E.g. verbing in the sentence "The verbing of nouns".
Q2 Highly unusual words that occasionally occur in conversational speech.
Q1 Words to avoid: non words, contain illegal characters or simply wrong, unpronounceable or alternative spellings of existing words.

Q1 exists because training data contains significant numbers of these words. The recognition word list is derived from words of the development text for which pronunciations are available and have a quality of Q2 or greater. The list is then augmented to yield 50,000 words by taking the most frequent words in all our LM corpora for which firstly pronunciations are available and secondly the associated quality is Q2 or higher. To facilitate this process user interface tools were developed and it is planned to make them publicly available. In 2007 a

Table 2. Parameter settings for the IHM segmenter for the RT'06 and RT'07 systems.

Hyper- parameter	P(speech)	Minimum duration (ms)	Insertion penalty	Silence collar (ms)
RT'06	0.25	500	-15	100
RT'07	0.25	200	-40	200

Figure 2. Histograms of (a) speech (b) silence segment durations for manual segmentation, forced alignment and the RT'06 and RT'07 configurations.

total of 1500 new pronunciations were added to account for the inclusion of the Fisher corpus [**?**](See Section 3.4) in training, plus an additional 1750 words to give full coverage of the AMI corpus [**?**] including partial words.

3.2 Front-end processing

IHM speech segmentation was based on the 2006 system [**?**]. There a multilayer perceptron (MLP) consisting of 15 input frames (of feature dimension 51), 50 hidden units and 2 output classes (speech/non-speech) was used that was trained on forced alignment derived segmentation of 150 meetings. Segmentation is performed using Viterbi decoding of scaled likelihoods from the MLP. This system uses several hyper-parameters: the speech/non-speech prior probabilities, segment minimum duration, segment insertion penalty and silence collar. Each of these can be tuned on development data. In previous years the class priors were chosen to minimise the frame error rate (FER). The remaining parameters showed no great influence on performance of the segmenter at the frame level but they still appear to have impact on recognition performance. Hence the remaining parameters were selected to provide a good match of duration histograms from manual segmentation and those obtained for automatically determined segmentation. Table 2 shows the hyper-parameters used in the RT'06 and RT'07 systems that were, obtained using the described approach. Figure 2 illustrates the difference between duration histograms. Note that the minimum duration constraint in RT'06 configuration caused a significant increase in segments of that length. The length of silence segments was also increased.

MDM front-end processing remained mostly identical to previous years. In contrast to before, if only two omni-directional microphones are present beam-

Table 3. %WER and %DER results on the *rt06seval* set using the ICSI RT'06 STT segmentation and several configurations of the TNO diarisation system. Fixed numbers speaker clusters disregard the actual number of people present.

		#clusters	WER	DER
ICSI	RT'06 eval	4	56.8	-
TNO	Optimised for DER	-	60.1	**18.1**
TNO	Fixed # clusters	6	56.2	30.9
TNO	Fixed # clusters	5	56.1	30.1
TNO	Fixed # clusters	4	**55.6**	33.6
TNO	Fixed # clusters	3	56.3	38.9
TNO	Fixed # clusters	1	56.9	64.0

forming is replaced by simple energy based switching. Secondly, in the case of speaker-directed directional no beam-forming is used either. Consistent gains with these simplifications were shown in [?]. In previous years the AMI systems made use of segmentation and speaker clustering information for MDM from ICSI/SRI. This year experiments with using the TNO diarisation system[?]. Table 3 shows WER and diarisation error rate (DER, see [?] for a definition) results using different configurations of the TNO diarisation system. The first line is the baseline performance of a two-pass adapted system with the ICSI segmentation as used in the AMI RT'06 system. Using a system that yields low DER however obtains considerably poorer WER results than a system that uses a fixed number of speaker clusters, which in turn almost doubles the DER.

3.3 New Training Data

Due to the recent interest in meetings several corpora are now available: Apart from the ICSI Meeting corpus [?] two phases of the NIST corpus [?], the ISL [?] recordings and the complete AMI corpus [?] contain manually transcribed meeting data. In addition, recordings from Virginia Tech University (VIT) and the Linguistic Data Consortium (LDC) are used in small quantities as test data in NIST evaluations. The main additions in 2007 were the completion of the AMI corpus and the second phase of the NIST corpus, thus adding a total of almost 70 hours of carefully transcribed meetings for training. The AMI corpus consists of 100 hours of meetings where 70 hours follow so-called scenarios where certain roles are acted by the meeting participants. The meetings are recorded at three different sites and due to the proximity to research a large percentage of non-native English speakers was present. Table 4 shows perplexity values of different language models, splitting the corpus along scenario, gender, and language of origin. It is clear that scenario meetings are far less complex than "normal" ones. A rather surprising effect is the difference between languages of origin. An investigation of out of vocabulary (OOV) words (not shown here) cannot explain the consider differences, with the speakers of French origin having lowest perplexity (even lower than native English speakers) and German speakers the highest.

Table 4. Perplexities of several LMs on the AMI corpus with distinctions on the basis of gender, meeting type and language of origin. LMs are constructed by interpolating the AMI corpus and the listed background material in 5-fold cross-validation.

Language models	Overall	male	female	Scenario	Non-Scen
Broadcast News	99.8	99.3	100.9	87.9	137.8
CTS	100.5	100.1	101.6	88.2	140.2
Meetings	102.7	101.6	105.4	91.2	138.8
Combined (inc Web-Data)	92.9	92.8	93.2	84.1	119.7

Language model	English	French	German	OtherEU	S. Asia	Rest of World
Broadcast News	105.2	**97.7**	**128.5**	113.3	112.0	102.8
CTS	105.9	**100.2**	**128.9**	114.4	115.0	104.0
Meetings	110.3	**98.0**	**126.8**	115.9	113.3	103.7
Combined (inc Web-Data)	96.9	**90.8**	**111.0**	103.0	104.7	94.9

Table 5. %WER results on rescoring of 4-gram *rt05seval* (NIST 2005 RT evaluation set) lattices. Models are trained on meeting data using the ML or MPE criteria.

Training		IHM		MDM	
	Iter	100h	170h	68h	130h
ML	-	24.0	23.6	39.7	38.1
MPE	1st	23.2	-	39.3	-
MPE	final	21.7	21.5	37.3	36.0

In RT'06 the aforementioned corpora amounted to approximately 100 hours of meeting data. The new additions bring the total to approximately 170 hours. The Tables 5 and 7 show the WER gains using the new and old data sets in direct comparison, with ML and MPE training, and for IHM and MDM data. Models are trained on meeting data only and a combination of PLP and LCRC features[**?**,**?**]. The MDM data set size is reduced from 170 hours to 130 hours due to the exclusion of overlapped speech (see [**?**]). Results in Table 5 seem to indicate that the overall gain from the additional data is moderate, at least in the case for IHM. Here both after ML and MPE training the difference in performance between models trained on almost twice as much data is modest. The gain for MDM is higher because the original set of data was very small.

A closer look at performance on AMI data in Table 7 however indicates a different picture. Here the difference for IHM seems to be around 2% WER and for MDM almost 3% WER absolute. In Table 7 the MPE results are compared for all meeting sources. It is clear that the benefit from the new corpora was on data from the same source, whereas performance even degraded on others. This is most likely caused by increased under-representation in the overall training data set.

Table 6. %WER results on rescoring of 4-gram *rt05seval* lattices using AMI corpus data only.

Training		IHM		MDM	
	Iter	100h	170h	68h	130h
ML	-	21.7	19.9	34.6	31.8
MPE	final	19.6	18.1	32.0	29.1

Table 7. Gain of additional training data in 2007. %WER results on rescoring of 4-gram *rt05seval* lattices.

		AMI	CMU	ICSI	NIST	VT
IHM	100h	19.6	20.1	18.3	26.9	24.3
	170h	18.1	19.9	19.2	26.2	24.7
Δ		*-1.5*	*-0.2*	*+0.9*	*-0.7*	*+0.4*
MDM	68h	32.0	27.2	35.7	37.9	45.3
	130h	29.1	27.8	34.0	35.5	45.3
Δ		-2.9	+0.6	-1.7	-2.4	+0.0

3.4 Training on 2000 hours of CTS data

In addition to incorporating new meeting data also the adapted model sets were improved. Previously CTS models were trained on the Switchboard and Callhome corpora and adapted to the meeting domain. This year we have included 2000 hours of Fisher corpus recordings [?]. The corpus data was prepared in the usual fashion, including the deletion of non-uniform amounts of silence at segment boundaries. A total of 170 hours of silence based on the manual segmentation was deleted. Table 8 shows results on the NIST CTS 2001 evaluation test sets using 270, 1000, and 2000 hours of training data respectively, where the 270 hour set is identical to the Cambridge University *h5train03* training set and does not include Fisher data. 2000 hour models ware initialised from 1000 hour ML trained models. Overall a 2.1% improvement in WER was observed.

In [?] and [?] we have presented a method to retain the benefit from wide-band data modelling while retaining the gain from CTS data adaptation. Unfortunately a detailed description cannot be given here and the interested reader is referred to those papers. Table 9 shows the performance on the *rt05seval* set using models trained in the mapped NB space on PLP features using VTLN, HLDA and MPE MAP. Note that the baseline performance for CTS models does not change for the 2000 hour models. However, after adaptation a 0.5% difference is observed. This is 1.3% lower than not adapting at all. Note the considerable performance differences across meeting rooms.

4 The 2007 AMIDA System

4.1 Acoustic Modelling

As in 2006 models were trained using either meeting data only or adapting from meeting data. The features used were either PLP features ore PLP features together with LCRC posterior features[?]. In addition to these new models were

Table 8. %WER on the 20001 NIST CTS evaluation set with different amounts of training data.

	#Iter	270h	1000h	2000h
ML	-	31.3	29.6	-
MPE	1	30.3	28.6	28.5
MPE	9	28.0	26.4	25.9

Table 9. %WER results on *rt06seval* adapting CTS models to meeting data including NB/WB transforms, joined HLDA, and MPE-MAP. 270/100 and 2000/100 refer to the amount of CTS and meeting data respectively.

270h / 100h	#Iter	TOT	AMI	CMU	ICSI	NIST	VT
CTS		30.4	31.4	33.0	26.4	32.5	28.3
MAPr-5iter	5	26.0	26.0	25.7	22.1	29.6	26.6
MPE-MAP	1	25.1	25.0	24.8	20.9	29.0	26.0
MPE-MAP	9	23.9	24.0	23.6	20.1	28.2	24.1

2000h / 170h	#Iter	TOT	AMI	CMU	ICSI	NIST	VT
CTS		30.4	30.7	31.3	27.9	32.2	30.1
MAP	6	23.8	22.8	23.0	20.8	27.1	25.5
MPE-MAP	1	23.2	22.1	22.4	20.5	26.3	25.2
MPE-MAP	7	22.1	20.4	20.2	19.7	25.7	24.8
No adaptation	-	23.4	20.5	21.2	20.2	29.0	26.6

also trained on Mel Frequency Cepstral Coefficients (MFCC) and an alternative posterior based feature vector, the Bottleneck (BN) features. BN features originate from a very similar process than LCRC features, however, instead of using the output of the MLPs the outputs of a hidden layer in a 5 layer network are used directly as features[**?**]. In detailed experiments these features were shown to yield approximately equivalent performance. Table 10 shows the performance of models trained on the *ihntrain07* training set using both feature representations and identical training style including speaker adaptive training. It can be observed that despite initially poorer performance the MFCC/BN based models yield almost identical error rates after discriminative training. For MDM only PLP/LCRC features were trained, with similar gains in each of the training stages as observed for the IHM models. The final equivalent result is 37.9% which is still substantially higher than the IHM performance.

4.2 Language Modelling

Language models (LMs) were constructed in a two-stage process. In the first instance out of more than 15 language models the nine most highly weighted LMs are selected and used as background language model for a web data search [**?**].

Table 10. Comparison of various front-end configurations. %WER Results on *rt06seval* using models trained on the 170 hour *ihmtrain07* training set.

Features	Tr	Adapt/Normalise	TOT	CMU	EDI	NIST	TNO	VT
PLP	ML		39.0	39.0	35.4	33.7	40.3	45.6
PLP	ML	VTLN HLDA	31.8	31.9	29.0	29.1	30.0	37.9
PLP + LCRC	ML	VTLN HLDA	-	-	-	-	-	-
PLP + LCRC	ML	VTLN HLDA SAT	27.2	27.2	25.0	25.0	27.1	32.1
PLP + LCRC	MPE	VTLN HLDA SAT	25.4	25.4	23.3	23.3	25.2	29.4

Features	Tr	Adapt/Normalise	TOT	CMU	EDI	NIST	TNO	VT
MFCC	ML		39.7	39.9	37.0	34.2	38.9	45.8
MFCC	ML	VTLN HLDA	34.2	34.2	32.6	29.9	32.0	41.0
MFCC + BN	ML	VTLN HLDA	29.4	29.3	27.5	26.6	28.1	35.6
MFCC + BN	ML	VTLN HLDA SAT	27.3	27.2	25.2	25.6	26.5	32.3
MFCC + BN	MPE	VTLN HLDA SAT	25.6	25.6	23.0	23.6	24.9	30.1

20MW of web data are collected and used to train an additional LM component. In the second stage a new LM is constructed from the ten most highly weighted LMs but components with a weight of less than 1% are removed. Table 11 shows the associated language model weights. Since web-data was already collected multiple times the newly collected data dropped out of the final list. The final LM had a perplexity of 73.1 on the *rt07seval* data.

As in previous years the AMIDA system was tested on lecture room data without training on any acoustic material from that domain. The only change was the training of a separate language model. In Table 12 perplexities of models trained and optimised for one domain are tested on the other one. It is clear that perplexities on conference style meetings are substantially lower in general. The models trained on lecture data appear to generalise better to conference room meetings than the reverse. This could be explained by the very generic nature of most conference recordings whereas the lecture room recordings have highly specialist content. Finally we have tested progress in language modelling for the past years. Table 13 shows WER results using the dictionaries and language models developed in each year (trigram). Single pass decoding was performed on the rt07seval data set. The results indicate small improvements each year even though the meeting sources changed considerably in those years.

4.3 System Architecture

Figure 3 shows an outline of the complete system. The system operates in a total of ten passes. However, not each pass does generate word level output. The output of a pass can either be word sequences, word lattices or confusion networks.

Table 11. LM interpolation weights for the two stages of LM construction. The left table shows the models used in the first stage, on the right are the models for the second stage.

corpus	weight
Fisher webdata from UW	0.220
AMI corpus eval	0.210
Fisher	0.186
Meetings webdata from UW	0.103
ISL meeting corpus	0.081
Switchboard Callhome	0.048
Swbd webdata from UW	0.045
AMI corpus webdata	0.038
Hub4 1996 LM	0.035
NIST meetings phase 2	0.029

corpus	weight
Stage 1 *conf LM*	0.912
rt06s conf webdata	0.054
ICSI meeting corpus	0.019
NIST meeting corpora	0.014

Table 12. Perplexities using LMs across lecture and conference room domains. *confmtg* and *lectmtg* denote the two domains as defined in NIST RT evaluations.

4-gram Models	*confmtg* (*rt06seval*)	*lectmtg* (*rt07slmdev*)
RT06 LM	75.2	125.8
confmtg STAGE1	73.2	144.5
confmtg STAGE2	73.1	140.8
lectmtg STAGE1	82.9	120.4
lectmtg STAGE2	81.9	119.3

The initial two passes are identical to the 2006 system and ensure adaptation using VTLN and CMLLR. Passes P3, P5 and P8 use PLP/LCRC features, passes P4 P6 use MFCC/BN features and passes P7 and P9 use PLP features only. The models for the PLP/LCRC and MFCC/BN processing stages where described in Section 4.1 whereas the models for P7 and P9 are those outlined in Section 3.4. P3, P4 and P7 generate lattices and the confusion network outputs of P5, P8 and P9 are combined using ROVER to yield the final system result.

4.4 STT Results

Tables 14 and 15 show the performance of the 2007 system on IHM and MDM data respectively. These results should be compared with results in Table 1. Overall a reduction of 1.9% in WER is observed on *rt06seval* IHM. On MDM the difference between the 2006 and 2007 system is 3.8% WER absolute. Note that while the overall performance numbers on the *rt06seval* and *rt07seval* IHM sets are very similar. The underlying results for each meeting corpus differ much more on *rt07seval*. The difficulty in the CMU data originates from lower recording quality. Similar to systems before, the result of the third pass is already very close to the results of the final passes. The gap between first and third pass has narrowed due to more training data and MPE trained models. P5 and P6 output

Table 13. %WER results on the NIST RT 2007 conference room meeting test data *(rt07seval)* using trigram LMs of AMI systems in past and the current years. Dictionaries, word lists and language models change, acoustic models are from RT'07.

LM Year	TOT	CMU	EDI	NIST	VT
2005	28.7	33.6	20.7	14.7	31.7
2006	28.6	34.1	20.2	14.4	31.5
2007	28.5	34.0	20.3	14.4	31.1

Figure 3. Outline of the system passes of the AMIDA 2007 RT system.

yield similar WERs but a combination of the two does not decrease WERs, most likely due to cross-adaptation.

In Table 15 a comparison is made between ICSI/SRI and TNO segmentation and speaker information. Similar performance is observed. However, a comparison with manual segments still shows a substantial difference. The difference between MDM and IHM results is still large with almost 10% WER absolute.

4.5 SASTT Results

As mentioned before speaker attributed STT was a new evaluation task in 2007. Systems results were obtained by using the standard MDM systems and attaching speaker labels using a diarisation system. In Table 16 we compare the results with the system as described in 3.2and the TNO and ICSI diarisation systems. It can be observed that with a very low diarisation error rate of 5.1% approximately 1% of loss in WER is obtained.

The MDM system was also used to transcribe the RT'07 lecture evaluation set (*rt07slecteval*). As in previous the acoustic models and all front-end processing was take from the conference domain. Only specific language models were trained (See Section 4.2). ICSI/SRI MDM segmentation optimised for speech recognition (not diarisation!) was used for both recognition and speaker assignment. From results shown on conference data it is clear that this is sub-optimal. On *rt07slecteval* the STT overall performance was 48.2% WER absolute while the SASTT score was 65.2%.

Table 14. %WER results of the AMIDA RT 2007 system on the IHM *rt06seval* and *rt07seval* data sets.

	rt06seval						*rt07seval*				
	Tot	CMU	EDI	NIST	TNO	VT	Tot	CMU	EDI	NIST	VT
P1	35.4	35.4	32.5	31.5	35.2	39.8	37.4	47.7	29.3	33.8	38.4
P3	24.9	24.9	23.0	22.4	25.0	29.3	28.2	37.9	21.9	24.6	27.9
P4	24.4	24.4	22.7	21.7	23.9	28.8	27.9	38.0	21.7	24.1	27.4
P5 CN	**23.4**	**23.4**	**21.6**	**20.8**	**24.0**	**27.8**	**25.9**	**35.1**	**20.4**	**21.8**	**25.7**
P6 CN	23.5	23.5	21.7	21.0	23.9	27.7	25.7	34.9	20.4	21.5	25.7
P7	24.1	24.0	22.8	22.2	22.4	28.7	27.9	36.7	23.1	24.2	27.2
P8 CN	22.9	22.9	21.1	20.7	22.5	27.3	25.4	34.5	20.4	21.1	25.3
P9 CN	23.7	23.6	22.4	21.9	22.2	27.9	26.3	35.3	22.3	21.8	25.4
P5 + P8 + P9	**22.3**	**22.2**	**20.7**	**20.2**	**22.1**	**26.7**	**24.9**	**33.9**	**19.8**	**20.9**	**24.7**

Table 15. %WER on *rt07seval* using MDM data.

	ICSI S&C				AMI/DA S&C			
	TOT	Sub	Del	Ins	TOT	Sub	Del	Ins
P1	44.2	25.6	14.9	3.8	44.7	25.7	16.3	2.7
P3	38.9	18.5	16.8	3.5	34.5	19.3	12.5	2.7
FINAL	33.7	20.1	10.7	2.9	33.8	19.2	12.2	2.4
FINAL manual seg	30.2	18.7	9.4	2.0	-	-	-	-

5 Conclusions

We have presented the 2007 AMIDA system for the transcription of meetings and have shown results on the latest evaluation test sets. Major improvements in performance come from new data, fine-tuning of system parameters and a consolidation of system building processes.

Acknowledgements

This work was largely supported by the European Union 6th FWP IST Integrated Projects AMI and AMIDA (Augmented Multi-party Interaction / with Distant Access). We also would like to thank Andreas Stolcke and Chuck Wooters from the ICSI/SRI team for providing the segments and speaker labels and their diarisation output for MDM data.

References

1. Fiscus, J.: Spring 2007 (RT-07) Rich Transcription Meeting Recognition Evaluation Plan. U.S. NIST. (2007)
2. Hain, T., Burget, L., Dines, J., Garau, G., Karafiat, M., Lincoln, M., McCowan, I., Moore, D., Wan, V., Ordelman, R., Renals, S.: The development of the AMI system for the transcription of speech in meetings. In: Proc. MLMI'05. (2005)
3. Hain, T., Burget, L., Dines, J., Garau, G., Karafiat, M., Lincoln, M., McCowan, I., Moore, D., Wan, V., Ordelman, R., Renals, S.: The 2005 AMI system for the transcription of speech in meetings. In: Proc. NIST RT'05, Edinburgh (2005)

Table 16. %WER, and % speech activity detection (SAD) and % diarisation error rate (DER) results on MDM *rt07seval* using SASTT scoring. SpSub denotes the percentage of cases where the wrong speaker label was assigned to the correct word.

	%SAD	%DER	Sub	SpSub	Del	Ins	TOT
ASR optimised	-	-	19.3	9.5	12.1	2.3	43.2
TNO 2007 diarisation system	6.7	18.9	19.2	3.9	12.1	2.4	37.6
ICSI 2007 diarisation system	3.3	5.1	19.1	0.9	12.3	2.3	34.7

4. Fitt, S.: Documentation and user guide to UNISYN lexicon and post-lexical rules. Technical report, Centre for Speech Technology Research, Edinburgh (2000)
5. Burget, L.: Combination of speech features using smoothed heteroscedastic linear discriminant analysis. In: Proc. ICSLP, Jeju Island, Korea (2004) 4–7
6. Povey, D.: Discriminative Training for Large Vocabulary Speech, Recognition. PhD thesis, Cambridge University (2004)
7. Gales, M.J., Woodland, P.: Mean and variance adaptation within the mllr framework. Computer Speech & Language **10** (1996) 249–264
8. Hain, T., Burget, L., Dines, J., Garau, G., Karafiat, M., Lincoln, M., Vepa, J., Wan, V.: The ami meeting transcription system : Progress and performance. In: Proc. NIST RT'06 Workshop. Springer LNCS (2006)
9. Cieri, C., Miller, D., Walker, K.: The fisher corpus: a resource for the next generations of speech-to-text. In: LREC 2004: Fourth International Conference on Language Resources and Evaluatio, Lisbon (2004)
10. Carletta, J., Ashby, S., Bourban, S., Guillemot, M., Kronenthal, M., Lathoud, G., Lincoln, M., McCowan, I., Hain, T., Kraaij, W., Post, W., Kadlec, J., Wellner, P., Flynn, M., Reidsma, D.: The AMI meeting corpus. In: Proc. MLMI'05, Edinburgh (2005)
11. van Leeuwen, D.A., Huijbregts, M.: The ami speaker diarization system for nist rt06s meeting data. In: Proc. MLMI 2006. (2006) 371–384
12. Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C.: The ICSI meeting corpus. In: Proceedings IEEE ICASSP. (2003)
13. Garofolo, J., Laprun, C., Miche, M., Stanford, V., Tabassi, E.: The nist meeting room pilot corpus. In: Proc. LREC 2004. (2004)
14. Burger, S., MacLaren, V., Yu, H.: The ISL meeting corpus: The impact of meeting type on speech style. In: Proc. ICSLP. (2002)
15. Schwarz, P., Matějka, P., Černocký, J.: Hierarchical structures of neural networks for phoneme recognition. In: Accepted to IEEE ICASSP. (2006)
16. Karafiat, M., Burget, L., Hain, T., Cernocky, J.: Application of cmllr in narrow band wide band adapted systems. In: Proc 8th international conference INTERSPEECH 2007, Antwerp (2007) 4
17. Grezl, F., Karafiat, M., Kontar, S., Cernocky, J.: Probabilistic and bottle-neck features for lvcsr of meetings. In: Proc. ICASSP. Volume 4. (2007) IV–757–IV–760
18. Wan, V., Hain, T.: Strategies for language model web-data collection. In: Proc. ICASSP'06. Number SLP-P17.11 (2006)