The AMI Meeting Transcription System : Progress and Performance

Thomas Hain[1], Lukas Burget[2], John Dines[3], Giulia Garau[4], Martin Karafiat[2], Mike Lincoln[4], Jithendra Vepa[3], 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.
th@dcs.shef.ac.uk

Abstract We present the AMI 2006 system for the transcription of speech in meetings. The system was jointly developed by multiple sites on the basis of the 2005 system for participation in the NIST RT'05 evaluations. The paper describes major developments such as improvements in automatic segmentation, cross-domain model adaptation, inclusion of MLP based features, improvements in decoding, language modelling and vocal tract length normalisation, the use of a new decoder, and a new system architecture. This is followed by a comprehensive description of the final system and its performance in the NIST RT'06s evaluations. In comparison to the previous year word error rate results on the individual headset microphone task were reduced by 20% relative.

1 Introduction

Conference room meetings are an integral basis of business life. For many they constitute a main part of their daily work. Nevertheless meetings are often viewed as ineffective, hence many attempts are made to increase effectiveness while ensuring good communication. Recordings of meetings themselves are likely to be of little help. Instead the analysis of meeting content can be used for design tools that preserve the essential information in accessible form. The foundation for such analysis is in many cases the spoken word, hence work on meeting transcription is essential for the AMI project [5]. The transcription system presented in this paper is developed by multiple sites involved in the AMI project [1].

High degrees of variability present in the meetings recordings make it an interesting task for automatic speech recognition[2]. The speaking style is conversational by nature but the presence of multiple conversation partners results in characteristic speaking style. It was found that surprisingly Broadcast News (BN) material fits reasonably well (e.g. [3]). The diversity of topics appears to be large, however analysis of existing corpora is ambiguous[2]. Another obvious

[5] http://www.amiproject.org

source of variability is the recording conditions. The AMI system has focused on two conditions: the individual headset microphone (IHM) and the multiple distant microphone (MDM) conditions. While the latter seems to represent a natural situation, the former allows the establishment of baselines and assessment of the loss due to different recording setups.

In 2005 we presented our first system for participation in the NIST RT 2005 evaluations (Sys05)[4]. This initial system achieved state-of-the-art competitive performance both on conference and lecture room tasks and the system formed the basis of our development this year. Analysis of the system exhibited several issues. For example the difference in word error rate (WER) performance between manual and automatic segmentation was more than 20% relative on IHM while the difference between IHM and MDM results was approximately 30% relative. The latter was dependent on the recording setup with generally larger differences where the setup was less strictly specified. Other less prominent issues were addressed in this paper, such as speed, stability of vocal tract length normalisation (VTLN), pronunciations, adaptation of CTS models, etc.

In the following section we briefly outline the main characteristics of the 2005 system, followed by a section discussing experiments and algorithmic differences for various components in the 2006 system. This is followed by a section describing the final system architecture and results on conference and lecture room tasks. The final section concludes the paper.

2 The AMI 2005 STT System

The AMI 2005 STT system operates in a total of six passes[4][6]. The system is identical in structure both for IHM and MDM input. The systems differ in the front-ends and the acoustic models. Hence we focus initially on the description of the IHM system and highlight the differences for MDM later on.

The IHM front-end converts the recordings into feature streams, with vectors comprised of 12 MF-PLP features and raw log energy and first and second order derivatives are added. The audio stream is split into meaningful segments. The segmenter uses echo cancellation prior to classification with a multi-layer perception (MLP). After segmentation cepstral mean and variance normalisation (CMN/CVN) is performed on a per channel basis (see Fig.1).

The first decoding pass yields initial transcripts that are subsequently used for estimation of VTLN warp factors. The feature vectors and CMN and CVN are recomputed. The second pass processes the new features and its output is used to adapt models with maximum likelihood linear regression (MLLR). In the third pass word lattices are produced which are rescored with trigram language models (LMs) and meeting room specific 4-gram LMs in the fourth pass. In the fifth pass acoustic rescoring with pronunciation probabilities is performed and the lattices are compressed into confusion networks (CNs) in the final pass. Acoustic models are trained on the *ihmtrain05* training set which merges four meeting corpora (the NIST, ISL, ICSI corpora and a preliminary part of the AMI

[6] Appropriate references to well known techniques mentioned in this section can be found in this paper.

	TOT	Sub	Del	Ins	Fem	Male	AMI	ISL	ICSI	NIST	VT
IHM	30.6	14.7	12.5	3.4	30.6	25.9	30.9	24.6	30.7	37.9	28.9
MDM	42.0	25.5	13.0	3.5	42.0	42.0	35.1	37.1	38.4	41.5	51.1

Table 1. Final results with the 2005 system on the *rt05seval* test set.

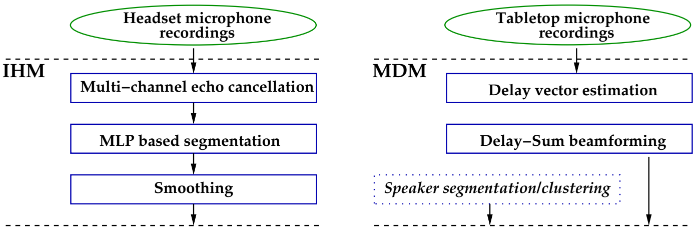

Headset microphone recordings
IHM
Multi–channel echo cancellation
MLP based segmentation
Smoothing
Tabletop microphone recordings
MDM
Delay vector estimation
Delay–Sum beamforming
Speaker segmentation/clustering

Figure 1. Front-ends for both IHM and MDM conditions

corpus). Model training included discriminative training and a smoothed version of heteroscedastic linear discriminant analysis (HLDA). Bigram, trigram and 4-gram language models are trained on a large variety of texts and specifically collected data by harvesting the word wide web.

The difference between MDM and IHM lies in the front-end and the acoustic model training set. The front-end operates in four stages: initial gain calibration is followed by noise compensation and frame based delay estimation between channels. The delay estimates are then used in superdirective beam-forming to yield a single output channel. All further steps were similar to the IHM case, segmentation and speaker clustering information for the MDM system were kindly provided by SRI/ICSI[3]. We repeat the results on the NIST RT 2005 conference room evaluation set (*rt05seval*) for convenience in Table 1. The system operated in 200-300 times real-time.

3 New developments in the 2006 system

In the 2005 system we could identify a series of major and minor weaknesses of the system of which some were addressed. Further, as the 2005 system was our initial move not all components had been developed as far as we would have liked and hence we also continued on this path to include new technologies. The main sets of changes to the system include: Improved segmentation for IHM; standard unsmoothed HLDA with removal of silence; posterior probability based features [5]; speaker adaptive training (SAT) with constrained MLLR (CMLLR) [6]; acoustic feature space mappings and maximum-a-posteriori (MAP) adapted HLDA; search model based LM data collection; as well as a modified system architecture that includes the use of a new decoder. In the following sections we present more details on these changes.

3.1 Improved Front-ends

Several changes to both the IHM and MDM front-ends (see Figure 1) were made.

Segmentation	TOT	EDI	TNO	CMU	VIT	NIS
manual	40.4	32.8	45.4	43.4	41.6	40.6
automatic	41.4	33.7	45.9	43.4	43.6	42.5

Table 2. WER on *rt06seval IHM* first passes using manual and automatic segmentation.

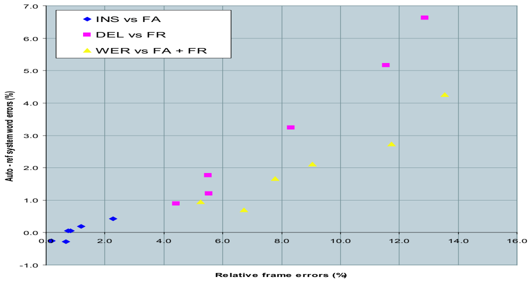
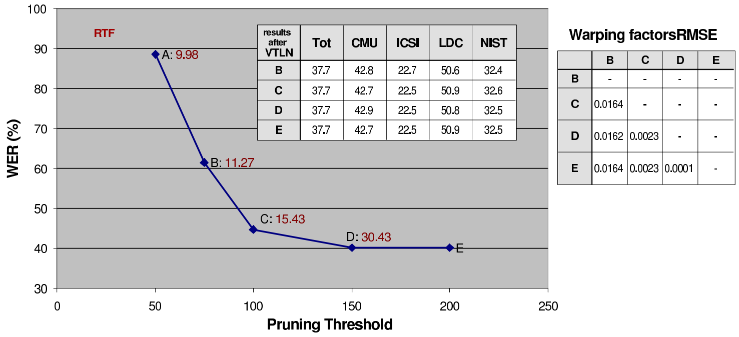


Figure 2. False Alarm (FA) and False Reject (FR) frame error rate changes in relationship to WER changes between manual and automatic segmentation on *rt06seval*.

Individual Headset Microphone The initial cross-talk suppression is based on an adaptive LMS echo canceller [7] followed by MF-PLP feature extraction. Different to last year, features to aid in the detection of cross-talk are extracted from the original recording (prior to cross-talk suppression). These features are cross-channel normalised energy, signal kurtosis, mean cross-correlation and maximum normalised cross-correlation. The cross-channel normalised energy is calculated as the energy for the present channel divided by the sum of energies across all channels [8]. In addition the MLP setup was changed to include 50 hidden units and the models are trained on 90 hours of data from all meetings in the *ihmtrain05* set. On *rt05seval* the first pass was found to give almost identical results to manual segmentation. Table 2 shows a comparison of manual versus automatic segmentation on *rt06seval*.

Figure 2 shows the correlation between WER and frame error rates for the meetings in *rt06seval*. The sum of false alarm (FA) and false reject (FR) rates exhibit a linear relationship with word errors. The main contributor are FR errors, which are unrecoverable.

Multiple Distant Microphones Only minor changes were made. Analysis on *rt05seval* showed that the system performed poorly on recordings from the VT meeting room. The reason was the use of only two microphones that were placed far apart in the room, causing delay estimation and hence poor beam-forming. The solution was to simply pick the channel with the highest energy for every time frame. This approach was also beneficial for the *rt06seval* set where the VT recording setup included four microphones directed at the speakers. Further problems had been caused by mis-aligned audio files, a problem eliminated in our 2006 system. Overall these changes brought improvements of 2.2% WER absolute on *rt05seval* in the first pass, and a 6% absolute change on VT data.



results after VTLN	Tot	CMU	ICSI	LDC	NIST
B	37.7	42.8	22.7	50.6	32.4
C	37.7	42.7	22.5	50.9	32.6
D	37.7	42.9	22.5	50.8	32.5
E	37.7	42.7	22.5	50.9	32.5

Warping factorsRMSE

	B	C	D	E
B	-	-	-	-
C	0.0164	-	-	-
D	0.0162	0.0023	-	-
E	0.0164	0.0023	0.0001	-

Figure 3. WER results on *rt04seval IHM* in the second pass. Real time factors (RTFs) combine first and second pass. The table shows RMSE results for operating points in the first and second pass. RMSE denotes the root mean squared warp factor difference to the baseline system.

SAT iterations	PLP	PLP+LCRC
-	28.7	25.2
adapt	27.9	24.2
1	27.6	24.1
2	27.4	24.0

Table 3. WER results for SAT rescoring 4-gram lattices on *rt05seval* IHM. LCRC denotes posterior based features

3.2 Vocal tract length normalisation experiments

Maximum likelihood based VTLN was part of the 2005 system, where relative WER improvements of more than 10% for both IHM and MDM were found. However, the cost in terms of complexity and real time was large, since the first pass is only devoted to finding initial transcripts for VTLN. Experiments were conducted to determine the importance of high quality transcripts. Fig. 3 shows WER results in the second pass as a function of real time factors and associated pruning in the first pass. On the right side the effect of pruning in both passes on the warp factor estimates compared to the baseline is shown. The operating point C was chosen for the first pass and D for the second pass.

3.3 Speaker adaptive training

The system already makes use of multiple speaker and channel normalisation techniques. Both CMN/CVN and VTLN yield substantial gains in WER close to 20% relative. Both techniques are simple and have few parameters. Speaker adaptive training (SAT) allows further normalisation and was already successfully applied in [3]. Here constrained MLLR[6] with two transforms was used where one is designated to the silence models.

3.4 Posterior based features

MLP based features have been deployed in large ASR systems (e.g. [9]). In [5] a similar approach was taken to produce a set of posterior probability based fea-

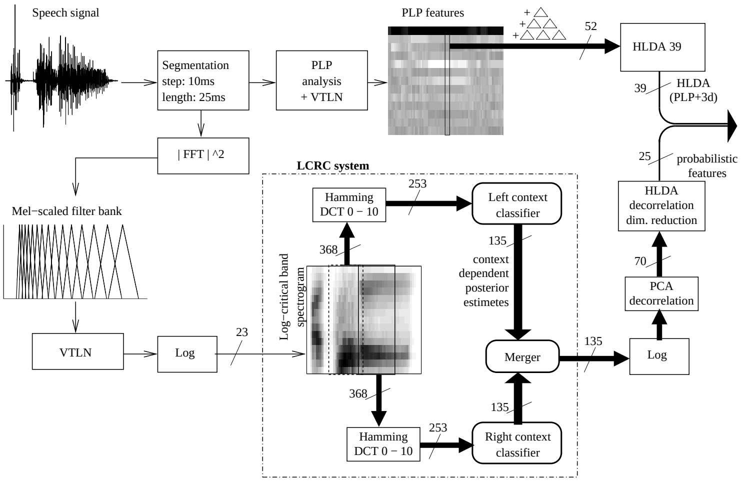


Figure 4. Computation of LCRC features.

tures computed with multiple layers of MLPs. While these kinds of features do not yield good performance directly, they clearly hold complementary information to the standard PLPs or MFCCs, and can bring substantial improvements in WER. Figure 4 describes the creation process of the feature vector. The top part shows standard MF-PLP generation and projection into a 39 dimensional space using HLDA. For the generation of the LCRC features first standard VTLN and CMN/CVN is applied to Mel frequency log filterbank (FB) coefficients. 23 FB coefficients are extracted every 10ms and 15 vectors of left context are then used to find the LC state level phone posterior estimates. The same procedure is performed with the right context. These posteriors are then combined with a third MLP network and after logarithmic compression the 135 dimensional feature vector is reduced to dimension 70 using principal component analysis. This step is only necessary because the final dimensionality reduction using HLDA was not feasible with such high dimensional vectors. The final 25-dimensional feature vector is appended to the standard 39 dimensional feature. Mean and variance normalisation is repeated at this stage.

Table 4 shows results for different combinations of training strategies and feature vectors. The number of states and mixture components remained constant and all systems use VTLN. Despite a considerable increase in dimensionality and data sparsity substantial performance improvements was found. The gains appear to be independent of the underlying training strategies. Note that the results in Table 4 may be somewhat biased because they were obtained by lattice rescoring.

3.5 Adapting to the meeting domain

One of the main short-comings of the 2005 system was the fact that only meeting data (*ihmtrain05/mdmtrin05*) could be used for discriminative training. Both

System	Training critrion	PLP	LCRC+PLP
Baseline	ML	28.7	25.2
SAT	ML	27.6	23.9
SAT	MPE	24.5	21.7

Table 4. WER results on *rt05seval*/IHM rescoring Sys05 4-gram lattices. Contrasting LCRC features using SAT and MPE

sets are comparatively smaller than the CTS training set and experimental evidence suggests that discriminative training should perform better ([10]) with more data. However, CTS and meeting data have different bandwidth and initial experiments showed that joint adaptation and projection into common space yields better performance[1]. The use of HLDA complicates matters considerably since it is not clear in which domain the matrix should be trained. It was decided to project the meeting data into the narrowband space where both HLDA statistics can be gathered and discriminative training be performed without regeneration of training lattices.

Initial full covariance statistic is estimated on the CTS training set. A single CMLLR transform is trained to map the 52D wideband (WB) meeting data to a 52D narrowband (NB) CTS space. The meeting data is mapped with this transform and full covariance statistics is obtained using models based on CTS phonetic decision tree clustering. The two sets of statistics are combined with MAP-like equations. The combined set of statistics is used to obtain a joint HLDA transform (JT). Now combined models in JT space can be trained using both CTS and mapped meeting data. These are then used to retrain CTS models in JT space, followed by speaker adaptive training and minimum phone error (MPE) training[10]. Equivalently to adaptation of maximum likelihood models with MAP, the JT/SAT/MPE models are adapted to meeting data using MPE-MAP[11]. The inclusion of SAT requires the presence of transforms on meeting data. These are obtained from SAT training of MAP adapted CTS models in JT space. Overall the performance improvement of this procedure was at least 0.6% on rt06seval. However, the elaborate process prohibited inclusion of LCRC features at this point.

3.6 Web data collection

Language model data for meetings are a rare resource and hence all sites have included language model material collected from the world wide web using approaches as originally described by [12]. The technique is based on sending a set of queries to well known Internet search engines and harvesting of the resulting documents. In Sys05 we used data collected using only queries (n-grams) that were not already present in our background LM material. Since then we have refined this approach [13]. In this work we use search models to predict the benefit of the search results on perplexity. The set of N-grams (word w with history h) present in a sample text T is ranked inversely with

$$\sum_v \frac{(\alpha P(w|h,T) + \beta P(w|h,B))}{(\alpha P(v|h,T) + \beta P(v|h,B))}$$

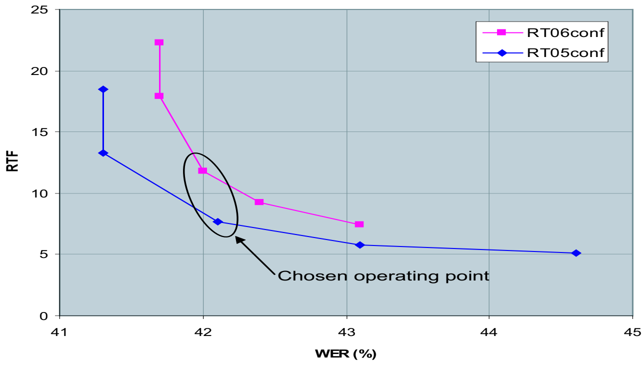


Figure 5. WER results on IHM data with the first pass using Juicer.

The probability estimates $P(w|h, T)$and $P(w|h, B)$ are provided by language models trained on the sample text and the background material B. The weights α and β were set equal. Small gains in perplexity were found with moderate data set sizes.

3.7 Juicer

As already discussed above, the 2005 system had a high RTF, partly due to slow initial stages. A second approach to address this is a faster decoder. Juicer [14] is a large vocabulary speech decoder based on weighted finite-state transducer (WFST). It uses a time-synchronous Viterbi search based on the token-passing algorithm with beam-search and histogram pruning. Juicer works with a single WFST composed of language model, dictionary and acoustic model. For the composition and optimisation of WFST resources, Juicer relies on the functionality of the AT&T finite-state machine library [15] and MIT FST toolkit [16]. The main advantage of WFST-based decoders is the decoupling of the decoding network generation and the actual decoding process. But there are limitations in composing the decoding networks, mainly due to high memory requirements, when used with large higher-order N-gram language models. Hence, pruned trigram language models were used for constructing decoding networks. Figure 5 shows performance versus RTF. The overall performance is within 1% absolute of the best results with HDecode[7].

4 System Architecture

Figure 6 shows the 2006 system architecture in diagrammatic form. In comparison to Sys05 the following major changes were made: The initial pass P1 now includes posterior feature computation; the output of P1 is used for both VTLN and adaptation of SAT models in pass P2. Lattice generation is performed in P2, with lattice expansion to unified 4-gram lattices in P3. These lattices are then rescored with different acoustic models. The original plan was to perform system combination by combining confusion networks, however this turned out

[7] HDecode is distributed as part of HTK (`http://htk.eng.cam.ac.uk`)

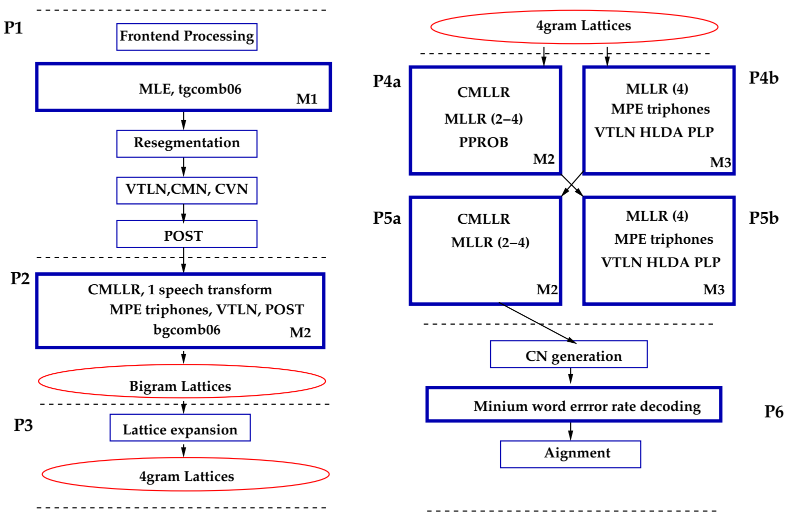


Figure 6. Processing steps of the AMI 2006 system.

to yield poorer performance. The best performing path was P4b followed by P5a. These passes are similar to the lattice rescoring passes in sys05, however include standard MLLR adaptation on top of the use of constrained MLLR.

5 System Components

Most of the system software is based on HTK. In particular three different decoders were included: In P1 Juicer (see Section 3.7) is used; in P2 HDecode provides lattice generation capability; passes P4 and P5 operate with HVite for rescoring. Confusion networks are generated using the SRI LM toolkit.

5.1 Acoustic Models

All acoustic models are phonetically state tied mixture of Gaussian HMMs with 16 mixture components. For the IHM system models were trained on *ihmtrain05*, for MDM models were trained on *mdmtrain05*. LCRC MLP models are trained on 30 hour subsets. The models M1 (see Figure 6) are identical to those used in Sys05. Models M2 are trained on PLP+LCRC features using SAT and MPE as outlined above. Several iterations of SAT were necessary for improved performance, followed by a total of 15 iterations of MPE training. For IHM models both MMI and MPE numerator and denominator statistics were combined with fixed weights. The M3 models are trained in the form outlined in Section 3.5, on the 300 hour *h5train03* training set[4].

5.2 Vocabulary, Language Models and Dictionaries

The vocabulary was built in similar fashion to Sys05 and changed only moderately. New web-data was collected for both conference and lecture room meetings with the technique outlined in Section 3.6. Table 5 shows perplexity results on

Perplexity	conference			lecture		
	2g	3g	4g	2g	3g	4g
2006 LM	106.9	86.2	82.7	157.9	127.6	122.4
2005 LM	105.6	84.3	81.2	165.6	137.4	134.5

Table 5. Language model perplexity results on *rt05seval*.

	TOT	Sub	Del	Ins	Fem	Male	AMI	CMU	ICSI	NIST	VT
IHM	23.7	12.0	9.9	1.7	23.7	20.3	22.0	20.1	21.1	30.0	25.7
MDM	33.0	18.7	12.3	2.1	33.0	35.4	28.8	32.6	35.8	35.4	33.7

Table 6. WER results of Sys06 on *rt05seval*.

rt05seval (both conference and lecture room meetings). Note that the 2005 language models shows lower perplexity. The reason for this behaviour is that the 2006 LM interpolation weight estimation did not include ICSI data as it was not part of *rt06seval*. The new method of data collection appeared to work well on lecture room data.

6 Overall System Performance

The development performance of the AMI 2006 system (Sys06) on the *rt05seval* data set is shown in Table 6 and can be directly compared with the results shown in Table 1. It is clear that at least on this test set substantial improvements have been made. The main improvements of the IHM system appear on the AMI data, while MDM improvement is highest on the VT subset.

Table 7 shows IHM results on the 2006 evaluation set, both with automatic and manual segmentation. The huge difference between initial and final pass results is even larger than before due to faster processing. After the third pass the results are already very close to the final performance, especially for manual segmentation. Even though the P4b system has lower performance on its own the inclusion into the adaptation path yields a further 0.5% absolute. Simple adaptation with P4a supervision did not give any improvement. It is interesting to note that automatic and manual segmentation differ minimally initially however the difference is immediately obvious once the systems use adaptation. Since this cannot be a speaker labelling problem it is likely that this is caused by cutting into sentences.

The MDM performance is given in Table 8 (non-overlap results). Again the initial pass yields very poor performance and the difference between the output of the third pass and the final result is small. Overall the gap between IHM and MDM performance appears to be wider than on the *rt05seval* test set.

6.1 Lecture Room Meetings

Similar to Sys05, the only component changed to the conference room meeting transcription system was the language model. Neither dictionary nor any acoustic models were modified. Results for both IHM and MDM can be found in Tables 9 and 8 respectively. Note the poor performance of the initial pass of both systems. In the IHM case however the system recovers reasonably well. Substantial difference in WER between data sources is visible. Further the high

	Automatic Segmentation						Manual segmentation					
	TOT	CMU	EDI	NIST	TNO	VT	TOT	CMU	EDI	NIST	TNO	VT
P1	42.0	41.9	41.0	39.0	42.1	44.8	40.3	40.4	39.5	38.7	37.6	40.9
P2a	29.2	29.2	27.4	27.7	29.5	32.4	26.5	26.7	25.5	26.6	22.3	28.8
P3.tg	26.6	26.3	25.2	25.7	27.0	29.9	21.1	21.2	19.7	21.8	17.0	23.9
P3	26.0	25.7	24.6	25.2	26.3	29.5	22.9	22.9	22.3	23.8	19.0	25.1
P4a	25.1	25.0	22.8	23.8	26.0	29.1	21.9	21.9	20.7	22.6	18.1	24.6
P4b	25.6	25.3	23.8	24.9	24.3	29.8	22.5	22.5	21.8	23.6	17.2	25.6
P5a	24.6	24.4	22.6	23.6	24.1	28.8	21.5	21.5	20.3	22.4	17.1	24.2
P5a-cn	24.2	24.0	22.2	23.2	23.6	28.2	21.1	21.2	19.7	21.8	17.0	23.9

Table 7. WER results with Sys06 on rt06seval.

	Conference				Lecture			
	TOT	Sub	Del	Ins	TOT	Sub	Del	Ins
P1	58.2	35.8	16.7	5.7	70.7	46.0	16.3	8.5
P2a	45.6	26.4	15.1	4.1	60.0	31.6	23.6	4.9
P3	42.0	24.5	13.2	4.4	58.2	30.8	22.0	5.4
P4a	41.7	22.9	14.9	3.9	57.8	28.5	24.3	4.9
P5	40.9	22.2	15.3	3.5	56.1	28.2	23.9	4.0

Table 8. WER results for Sys06/MDM on the RT06 conference and lecture room test sets.

deletion rate for MDM is unusual and is not mirrored in the conference room data.

7 Conclusions

We have presented the changes made to the AMI 2005 system for transcription of speech in meetings. The main performance improvements originate for improved front-ends both in terms of segmentation and feature generation. As in 2005, there is still a large gap between performance on automatic and manual segmentation. This and the large difference between IHM and MDM results will require increased attention. We have also made improvements towards a faster system with real time factors below 100.

Acknowledgements

This work was largely supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811). We also would like to thank Andreas Stolcke and ICSI for providing the segments and speaker labels for MDM data.

References

1. 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)

	TOT	Sub	Del	Ins	AIT	IBM	ITC	UKA	UPC
P1	88.2	10.4	68.2	9.6	100.5	60.4	89.6	92.7	92.3
P2a	39.2	21.8	7.3	10.2	66.3	34.8	37.4	29.9	44.5
P3	37.1	20.1	6.6	10.4	63.6	31.6	35.6	28.6	41.6
P4a	35.1	19.2	6.9	9.0	56.5	31.5	33.5	27.5	39.9
P4b	37.2	20.5	6.7	10.1	66.1	32.0	36.5	28.5	39.8
P5a-cn	33.4	18.6	6.3	8.4	56.6	29.1	32.1	25.6	37.6

Table 9. WER results for Sys06/IHM on lecture room data.

2. Hain, T., John Dines and, G.G., Karafiat, M., Moore, D., Wan, V., Ordelman, R., Renals, S.: Transcription of conference room meetings: an investigation. In: Proc. Interspeech'05. (2005)
3. Stolcke, A., Anguera, X., Boakye, K., Cetin, O., Grezl, F., Janin, A., Manda, A., Peskin, B., Wooters, C., Zheng, J.: Further progress in meeting recognition: The icsi-sri spring 2005 speech-to-text evaluation system. In: Proc. NIST RT'05 Workshop. (2005)
4. 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 Workshop, Edinburgh (2005)
5. Schwarz, P., Matìjka, P., Cernocký, J.: Towards lower error rates in phoneme recognition. In: Proc. of 7th Intl. Conf. on Text, Speech and Dialogue. Number ISBN 3-540-23049-1 in Springer, Brno (2004) 8
6. Gales, M.J.F.: Linear transformations for hmm-based speech recognition. Technical Report CUED/F-INFENG/TR-291, Cambridge University Engineering Department (1997)
7. Messerschmitt, D., Hedberg, D., Cole, C., Haoui, A., P.Winship: Digital voice echo canceller with a tms32020. Application report SPRA129, Texas Instruments (1989)
8. Wrigley, S., Brown, G., Wan, V., Renals, S.: Speech and crosstalk detection in multichannel audio. IEEE Trans.Speech and Audio Processing **13** (2005) 84–91
9. Zhu, Q., Chen, A.S.B., Morgan, N.: Using MLP features in sri's conversationl speech recognition system. In: Proc. Interspeech'05. (2005)
10. Povey, D.: Discriminative Training for Large Vocabulary Speech, Recognition. PhD thesis, Cambridge University (2004)
11. Povey, D., Gales, M.J.F., Kim, D.Y., Woodland, P.C.: MMI-MAP and MPE-MAP for acoustic model adaptation. In: Proc. Eurospeech'03. (2003)
12. Bulyko, I., Ostendorf, M., Stolcke, A.: Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In: Proc. Human Language Technology Conference 2003. (2003)
13. Wan, V., Hain, T.: Strategies for language model web-data collection. In: Proc. ICASSP'06. Number SLP-P17.11 (2006)
14. Moore, D., Dines, J., Doss, M.M., Vepa, J., Cheng, O., Hain, T.: Juicer: A weighted finite state transducer speech decoder. In: Proc. MLMI'06. (2006)
15. Mohri, M., Pereira, F., Riley, M.: General-purpose finite-state machine software tools. Technical report, AT&T Labs -Research (1997)
16. Hetherington, L.: The mit fst toolkit. Technical report, L. Hetherington, "The MIT FST toolkit", MIT Computer Science and Artificial Intelligence Laboratory: http://people.csail.mit.edu/ilh/fst, May 2005. (2005)