The AMIDA 2009 Meeting Transcription System

Thomas Hain‡, Lukas Burget§, John Dines‡, Philip N. Garner‡, Asmaa El Hannani‡
Marijn Huijbregts‖, Martin Karafiat§, Mike Lincoln†,Vincent Wan‡

SpAndH‡	FIT§	Idiap‡	CSTR†	HMI‖
Univ. of Sheffield	Brno Univ. of Technology	Research Institute	Univ. of Edinburgh	Univ. Twente
Sheffield S1 4DP	Brno, 612 66	CH-1920 Martigny	Edinburgh EH8 9LW	7500 AE Enschede
United Kingdom	Czech Republic	Switzerland	United Kingdom	The Netherlands
th@dcs.shef.ac.uk	burget,karafiat@it.vutbr.cz	dines,pgarner@idiap.ch	m.lincoln@ed.ac.uk	marijn.huijbregts@let.ru.nl

Abstract

We present the AMIDA 2009 system for participation in the NIST RT'2009 STT evaluations. Systems for close-talking, far field and speaker attributed STT conditions are described. Improvements to our previous systems are: segmentation and diarisation; stacked bottle-neck posterior feature extraction; fMPE training of acoustic models; adaptation on complete meetings; improvements to WFST decoding; automatic optimisation of decoders and system graphs. Overall these changes gave a 6-13% relative reduction in word error rate while at the same time reducing the real-time factor by a factor of five and using considerably less data for acoustic model training.

Index Terms: speech recognition, meeting transcription

1. Introduction

Over the past 10 years the processing of meeting speech under a large variety of conditions and scenarios was the focus of many research groups. The progress made has attracted researchers to the field from outside the speech community, interested in higher level, downstream processing. With the advent of high quality telephone and video conferencing systems the opportunity to record, process, recognise, and categorise the interactions in meetings is recognised even by sceptics of speech and language processing technology. This area was also the focus of the AMI and AMIDA projects[1]: acquisition, multimodal recognition, and higher level processing of meetings, distributed and connected via teleconferencing or in a single room. Many components are necessary to capture interaction between people that does not require automatic speech recognition (ASR). However ASR naturally is the most important part to capture content.

While ASR is often solely associated with transcription, many applications in the meeting domain do not require full transcripts (e.g. content linking[2]). Nevertheless, formal evaluations conducted by the U.S. National Institute of Standards and Technology(NIST) focus on transcription. The system presented here was developed for the participation in the NIST RT'09 evaluations conducted in April 2009. These evaluations were the latest in a series that started in 2002, and where the AMI/AMIDA group participated since 2005[3]. Our previous contribution [4] has achieved very competitive performance for close talking conditions. In 2009 our main focus was on the far field condition where we have achieved the best result.

The test conditions in 2009 were similar to previous years. Data from different meeting rooms with a variety of recording configurations are processed. Two tasks are addressed: recording from individual head (IHM) as well as multiple distant microphones (MDM) in arbitrary configuration and number. The configuration can vary substantially by room , but configuration information for each room may be used. In 2009 one new aspect was added. Meetings from the AMIDA corpus, recorded on two sites, connected with video conferencing, were included.

The AMIDA meeting transcription system changed in many aspects: front-ends for both IHM and MDM, updated segmentation and a new MDM diarisation component; stacked bottleneck posterior features; fMPE training of acoustic models, adaptation on complete meetings; a substantially improved decoder; and local optimisation of system graphs. Overall this allowed an improvement between 6-13% relative reduction in word error rate (WER) while at the same time reducing the real-time factor by more than a factor of five. In the following sections we outline the changes in greater detail.

2. The AMIDA 2007 System

The AMIDA 2007 system[4] served as the base for development. The system accepts both IHM and MDM input and operates in several passes using cross-adaptation between passes using models with different front-ends, training strategies and training data. The key features in the system were: beamforming; MLP features; adaptation from 2000h of CTS retaining data using narrow band/wide band (NB/WB) transforms; speaker adaptive and minimum phone error (MPE) training; and decoding based on lattices. The real-time factor (RTF) on the IHM part was close to 100. In the following sections more detail on the system will be given where required. All comparisons in this paper are based on the NIST RT evaluation data sets for 2007 and 2009, denoted as *rt07seval* and *rt09seval* respectively.

3. Segmentation, Clustering and Filtering

The system development strategy is based on different enhancement, segmentation, and speaker labelling of IHM and MDM input but similar later processing stages.

3.1. Individual Head Microphone

Segmentation was performed in identical fashion to previous years[5]. A multi-layer perceptron based speech silence classifier is trained on MF-PLP and cross-talk features. Segmentation itself uses Hidden Markov Models (HMMs) for the purpose of setting duration constraints and speech/silence class priors, and an insertion penalty. For the 2009 system the models were retrained on parts of the 2009 IHM training set, naturally including the silence portions. More than 90 hours of audio and 290

	#Seg	Tot	CMU	EDI	NIST	VT
Ref	4527	29.3	36.7	24.5	24.5	31.2
30h	2717	32.6	41.2	26.2	29.1	33.3
90h	4541	31.7	42.4	25.3	26.8	31.7

Table 1: %WER on *rt07seval*. Ref is manual segmentation, and 30/90h give the amount of training data for MLPs. CMU/EDI/NIST/VT are meeting rooms.

Segmentation	Clustering	Unadapted	Adapted
Ref	-	42.1	36.3
auto	-	43.8	38.1
auto	Ref	40.1	31.1
auto	no delay	42.8	34.5
auto	with delay	42.1	32.7

Table 2: %WER on *rt07seval* using the first (unadapted) and third (adapted) pass of the RT'07 AMIDA MDM system

meetings were used for training. Table 1 shows results for a 2-pass adapted system on the *rt07seval* set. One can observe an overall gain of 0.9% WER absolute from retraining while the number of segments becomes closer to the reference. However, the gains are not uniform across meeting rooms. For lapel microphones (e.g. CMU) results get poorer, while for lower quality head microphones (NIST,VT) results improve substantially.

The results generalised reasonably to the *rt09seval* set as outlined in Sec. 6. However, one meeting from the NIST meeting room gave rise to a substantial WER differences between reference and automatic segmentation of more than 10% absolute. The reasons are likely to be imbalances in gain between microphone channels which are not automatically adjusted for. The segmentation is particularly vulnerable to this due to the use of cross channel energy features.

3.2. Multiple Distant Microphone

In previous years the output of a diarisation system developed by ICSI/SRI was used for segmentation and clustering. In 2009 that system was replaced by one based on [6], specifically adapted for ASR. Here the distant microphone channels are first Wiener filtered, followed by microphone array beamforming with the BeamFormIt toolkit[7]. The energy based beamformer delivers a single audio stream, together with relative delay estimates between channels. While in [6] only the beamformed audio was used for clustering, the delay values now augment standard MFCC features (for clustering only). Segments clustering is using the BIC criterion, with initial cluster number based on the amount of data. The MFCC and delay feature streams are normalised to yield identical average BIC scores.

Table 2 shows WER results of automatic approaches in comparison with the reference, for segmentation and speaker clustering. The loss from automatic segmentation alone is 1.7%, not surprisingly the difference after adaptation is similar. The difference between the unadapted results with or without speaker information originates from cepstral mean and variance normalisation (CMN/CVN) as that also is speaker based. Using delay features for clustering brings substantial performance gain, and the final loss from automatic speaker clustering is 1.6% WER absolute. Experiments indicate that the losses for automatic segmentation and clustering are almost additive.

3.3. Room filtering

One of the challenges for RT'09 were meetings held in two rooms. This implies that audio from room 1 was played through the loud-speaker in room 2. Thus speech from room 1 appears in the recordings in room 2. There is an unknown and variable audio transfer delay between rooms. Since the recording system and the transport video conferencing systems are independent echo cancellation was not possible. Performing segmentation in each room separately will yield different segments and it is unclear which are the correct ones. In order to filter out loud-speaker segments the following algorithm was used:

1. Take beam-formed audio file for each room

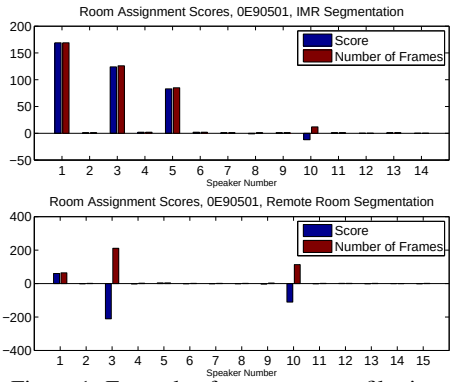

Figure 1: Example of room segment filtering

%Data retained	80%	90%	95%	100%
ML	42.6	42.2	42.8	42.8
MPE	40.7	40.5	40.7	40.8

Table 3: %WER on *rt07seval* using different thresholds on confidence scores in lattices.

2. Perform speaker segmentation on room 1 audio
   (a) For each speaker and frame, calculate the max. cross correlation between the audio from room 1 and room 2 (the delay). If delay > 0, increment room 1 count, otherwise the room 2 count
   (b) Assign speaker to room with highest count
   (c) Discard segments from speakers in room 2
3. Repeat using segmentation from room 2 audio, discarding segments assigned to room 1

Fig. 1 shows an example of frame counts. Speaker clustering output yielded too many clusters, but the important ones are clearly visible. With the above algorithm the single speaker in the remote room is clearly identified from each side. Results on the *rt09seval* multi-room meetings reveal that 3.1% WER absolute can be gained from using automatic room filtering compared to using only audio from one (i.e. the best) room. But naturally this number depends on the amount spoken in each room. Interestingly, the difference with reference segmentation is only 2.3% which seems to indicate that differences in segmentation are indeed a problem.

4. Modelling

4.1. Acoustic modelling

Several changes to acoustic modelling were made. The meeting training data originates from a variety of corpora (see [4]). In addition to the corpora used for the RT'07 system, about 6 hours of multi-room data was added (the AMIDA corpus), yielding a total of 177 hours of speech for IHM training. Data selection for MDM training however is not trivial: as training is performed on a single audio stream concurrent speech must be avoided. Automatic removal of overlapped speech is difficult. Removal of segments that contain any form of overlap would ignore more than 50% of the data. Hence automatic methods for finding and removing overlap need to be used. Based on alignment and word boundary times about 154 hours of training data were retained. However, alignment is often unreliable in boundary regions, even for IHM channels. An additional confidence based selection was used to remove an additional 10% of the data. Lattices were generated for the complete training set and ranked according to the highest word level posterior probability in the lattices. Table 3 shows results for maximum likelihood(ML) and MPE training. While reasonable gain is observed for ML,

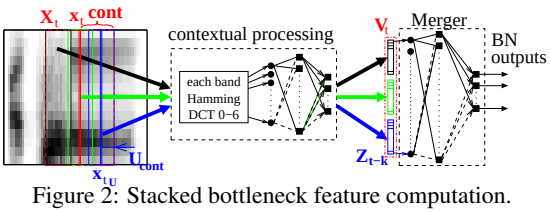

Figure 2: Stacked bottleneck feature computation.

HLDA-PLP	+BN	+LCRCBN	+SBN
36.0	31.7	30.6	29.4

Table 4: %WER on *rt07seval* using reference segmentation.

the impact on discriminative training (1 iteration) is modest. In addition to the changes in data, the BeamFormIt[7] beamformer was used. While conceptually identical to the system used before, the improved post-filtering allowed a reduction of 2.2%WER absolute on the *rt07seval* set.

In 2009 all models were trained on meeting data only. In that way the considerable complexity due to the use of the Fisher corpus for training was avoided. Naturally this comes at the cost of performance loss, as cross-adaptation with NB/WB models was shown to be very effective[4]. The simpler training setup allowed two changes. Use of two types of modified feature vectors, and fMPE training[8].

Bottle-neck (BN) features were introduced in [4] as a contrast to LCRC (left/right context) features. However, it is straight-forward to extend BN features with the LCRC paradigm. In this case the output of LC and RC BN MLPs forms the input to a 'merger' MLP, again with bottleneck output, resulting in LCRCBN features. Taking this concept further, the stacked BN features are presented in Fig. 2. Here the contextual MLP is shared between all contexts. This allows to reduce the number of parameters in the system. Table 4 shows a comparison of the feature types. Results are obtained using vocal tract length normalisation(VTLN) and CMN/CVN and BN features augment the PLP standard feature vectors. The resulting feature vector dimensionality ranges from 69 to 80. SBN features clearly outperform all other variants. However, for the purpose of complementarity LCRCBN features are also used.

fMPE is implemented using the RDLT framework[9]. Posterior probabilities of the Gaussian are computed for each frame and these are spliced with the averages of posteriors for adjacent frames 1-2, 3-5 and 6-9 on the right and likewise for the left context (i.e. 7 groups spanning 19 frames in total). All Gaussians in ML trained HMM model are pooled and clustered using agglomerative clustering to create a GMM with 1000 components. Only offset features (not the posteriors) are used. Table 5 shows results for use of fMPE in conjunction with BN features. As expected the gains are not additive and reduced with more complex time dependent features. Nevertheless an improvement of 1.4% WER can be observed.

4.2. Language modelling

Language model (LM) training data was kept identical to that used in 2007[4]. As OOV rates are generally found to be low using wordlist padding, only words from the 2007 evaluation data were added to the dictionary. However, two changes were triggered by using Juicer[10], a weighted finite state transducer

HLDA-PLP+	ML	MPE	fMPE	fMPE+MPE
-	35.6	32.6	31.4	29.7
+LCRCBN	30.4	28.1	26.7	26.3
+SBN	29.4	27.5	26.9	26.1

Table 5: %WER on *rt07seval*. Comparison of discriminative training and posterior features

Lexicon size	n–gram	Arcs in WFST	WER	RTF
2K	7	11.8M	55.3	0.827
6K	7	12.5M	48.2	0.625
10K	7	13.8M	47.2	0.582
16K	7	14.7M	46.8	0.589
50K	4	15.6M	46.8	0.579

Table 6: %WER results on *rt07seval* using different vocabulary size and n-gram order.

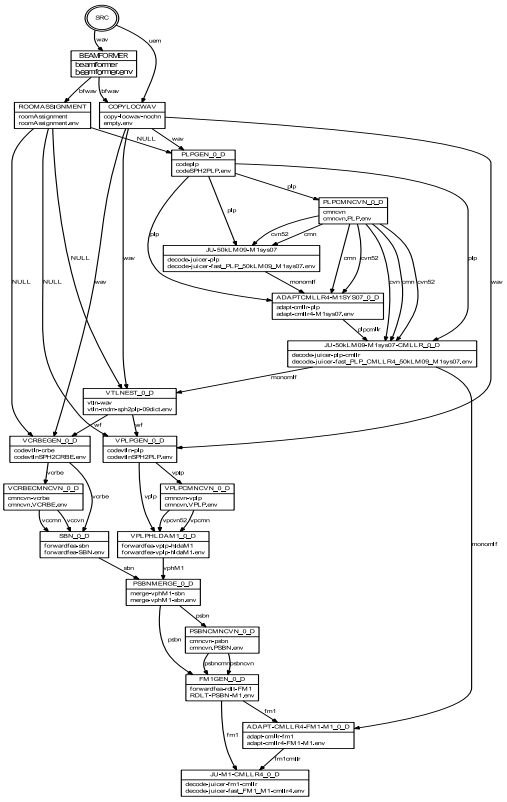

Figure 3: Graph of the MDM system

(WFST) decoder. Firstly, language model pruning is required to allow construction of WFSTs, which typically means a loss of 0.5%WER absolute. However, secondly the use of ngrams of higher order than 3 allows improved output. Table 6 shows results for language models of varying order and vocabulary size while approximately maintaining a certain decoder speed. The use of 7gram LMs with small vocabulary yields identical results to a 4gram LM with 50k vocabulary. This configuration allows for rapid initial adaptation (see Sec.5).

4.3. Decoding

Whereas the RT'07 relied mostly on HTK HDecode[4], the 2009 systems make almost exclusive use of Juicer[10], apart from lattice generation and rescoring. Juicer was considerably changed and is now substantially faster than HDecode at equal error rates, and as outlined before, also allows the use of higher order n-grams. Decoder parameters were optimised using methods outlined in [11] for each model configuration.

5. System overview and design

System design differed substantially from previous years. Instead of manual graph generation a semi-automatic approach was used. The resource optimisation toolkit (ROTK) allows the implementation of complex systems in the form of data

Description			Automatic				Manual			
LM	AM	Notes	**Tot**	**IDI**	**EDI**	**NIST**	**Tot**	**IDI**	**EDI**	**NIST**
6kLM09-7g	M2		41.3	45.1	32.3	44.9	38.3	44.0	31.9	38.3
50kLM09-4g	M1		45.9	50.9	36.8	48.3	43.7	50.2	36.8	43.3
50kLM09-4g	M2	CMLLR	36.4	38.8	28.5	40.2	32.9	37.9	27.7	32.5
50kLM09-4g	M3	CMLLR	28.3	28.5	21.4	33.2	24.2	27.8	21.1	23.5
50kLM09-4g	M4	Lattices / MLLR	27.6	28.3	20.9	31.9	23.9	27.9	20.6	22.8
50kLM09-4g	M3	Rescore / MLLR	27.2	28.0	20.3	31.9	23.5	27.5	20.0	22.6
Confusion network			27.4	28.6	20.4	31.6	23.8	28.0	20.7	22.5

Table 7: %WER on *rt09seval* IHM for the AMIDA 2009 system. IDI/EDI/NIST are meeting rooms.

Segmentation	Pass	*rt07seval*	*rt09eval*		
		Tot	Tot	Del	Ins
Automatic	First	40.3	44.2	10.8	4.7
	Final	29.3	33.2	9.3	3.2
Reference	First	37.8	42.3	10.3	3.2
	Final	26.5	30.7	8.3	2.1

Table 8: %WER on MDM for the AMIDA 2009 system.

processing graphs. Modules are for example: PLP computation; decoding using a specific configuration; adaptation; or segmentation. Once the modules are defined semi-automatic optimisation of graphs can be implemented. Thus the models and modules become more important than the exact processing sequence. For IHM the acoustic models developed were: HLDA-PLP/ML (M1) and HLDA-PLP/MPE (M2), VTLN/SBN/MPE/fMPE (M3), VTLN/LCRCBN/MPE/fMPE (M4). For MDM no LCRCBN models were created. The language models used are a 4g LM based on 50K vocabulary, and a 7gram LM with 6K vocabulary. For adaptation purposes a module for intersection of system output was added. Here, the intersection of two outputs, in terms of word and time, are retained. It was found in previous experiments on *rt07seval* that full meeting adaptation (rather than just on a 10 minute extract) yields improvements which are sustained or slightly improved when only adapting on intersection output (which typically discards half of the data).

Unfortunately an exhaustive search for all module combinations for identification of the best system is far too complex. Hence only local searches were conducted, as well as grid searches for locally optimal parameters. The result for MDM is given in Fig. 3. The IHM graph is considerably larger due to more models to choose from.

6. Results and Conclusions

The tables 7 and 8 show the overall performance. The IHM results for each decoding step are shown, for MDM only first and final passes are presented. The real-time factor (single thread) for the IHM system is 19.4, the output of M3 models is available at 9.84 RTF. The tables show the difference between automatic and manual segmentation. For IHM the large discrepancy in WER is mostly down to NIST data and the aforementioned issues with signal gain imbalance. The difference between automatic and manual segmentation on *rt09seval* is similar to that on *rt07seval* data. The WER difference between MDM and IHM on reference data is still high with 6.9% WER.

A wide range of new methods have been included into the AMIDA 2009 system for meeting transcription: : updated segmentation and MDM diarisation; stacked bottle-neck posterior features; fMPE training of acoustic models; adaptation on complete meetings; a substantially improved decoder; automatic optimisation of decoders and local optimisation of system graphs. Nevertheless, the system complexity overall was significantly reduced, in terms of training and amounts of training data, as well as RTF performance. The system can be tested on request by interested parties under `www.webasr.org`.

7. Acknowledgements

This work was partly supported by the European IST Programme Project AMIDA FP6-033812. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein. BUT researchers were partly supported by Grant Agency of Czech Republic Proj. No. 102/08/0707 and GP102/09/P635.

8. References

[1] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings: The AMI and AMIDA projects," in *Proc. ASRU 2007*, 2007, pp. 238–247.

[2] A. Popescu-Belis, J. Carletta, J. Kilgour, and P. Poller, "Accessing a large multimodal corpus using an automatic content linking device," in *Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*. Springer, 2009, vol. 5509, pp. 189–206.

[3] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals, "The 2005 AMI system for the transcription of speech in meetings," in *In Proc, MLMI*. Springer, 2005, vol. 3869, pp. 450–462.

[4] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, D. van Leeuwen, and V. Wan, "The 2007 AMI(DA) system for meeting transcription," in *Proc. MLMI*. Springer, 2007.

[5] J. Dines, J. Vepa, and T. Hain, "The segmentation of multichannel meeting recordings for automatic speech recognition," in *Proc. Interspeech 2006*, 2006.

[6] D. A. van Leeuwen and M. Huijbregts, "The AMI Speaker Diarization System for NIST RT06s Meeting Data," in *Proc. MLMI 2006*, 2006, pp. 371–384.

[7] X. Anguera, "Robust Speaker Diarization for Meetings," Ph.D. dissertation, UPC Barcelona, 2006.

[8] D. Povey, "Improvements to fMPE for discriminative training of features," in *Interspeech*, 2005, pp. 2977–2980.

[9] B. Zhang, S. Matsoukas, and R. Schwartz, "Recent progress on the discriminative region-dependent transform for speech feature extraction," in *Interspeech'06*, 2006.

[10] P. N. Garner, J. Dines, T. Hain, A. El Hannani, M. Karafiát, D. Korchagin, M. Lincoln, V. Wan, and L. Zhang, "Real-Time ASR from Meetings," in *Proc. Interspeech*, 2009.

[11] A. E. Hannani and T. Hain, "Automatic Optimisation of Speech Decoder Parameters," *IEEE Signal Processing Letters*, vol. 17(1), pp. 95 – 98, 2010.