THE AMI SYSTEM FOR THE TRANSCRIPTION OF SPEECH IN MEETINGS

Thomas Hain *Lukas Burget* *John Dines* *Giulia Garau*
Vincent Wan *Martin Karafiat* *Jithendra Vepa* *Mike Lincoln*

Department of Computer Science
Univ. of Sheffield
Sheffield S1 4DP
United Kingdom
{*th,v.wan@dcs.shef.ac.uk*}

Information Engineering Faculty
Brno Univ. of Technology
Brno, 612 66
Czech Republic
{*burget,karafiat*}*@fit.vutbr.cz*

IDIAP
Research Institute
CH-1920 Martigny
Switzerland
{*dines,vepa*}*@idiap.ch*

Centre for Speech Technology Research
Univ. of Edinburgh
Edinburgh EH8 9LW
United Kingdom
{*m.lincoln,g.garau*}*@ed.ac.uk*

ABSTRACT

This paper describes the AMI transcription system for speech in meetings developed in collaboration by five research groups. The system includes generic techniques such as discriminative and speaker adaptive training, vocal tract length normalisation, heteroscedastic linear discriminant analysis, maximum likelihood linear regression, and phone posterior based features, as well as techniques specifically designed for meeting data. These include segmentation and cross-talk suppression, beam-forming, domain adaptation, web-data collection, and channel adaptive training. The system was improved by more than 20% relative in word error rate compared to our previous system and was usd in the NIST RT'06 evaluations where it was found to yield competitive performance.

Index Terms— Speech recognition,, Meetings

1. INTRODUCTION

Many people spend considerable time in meetings despite of low complaints of low efficiency. So far computers are rarely used in streamlining the process and for extracting and retaining of the essential information. Projects like AMI (Augmented Multiparty Interaction) aim to investigate to use of machine based techniques to aid people in and outside of meetings to gain efficient access to information. Meetings are an audio-visual experience by nature, however verbal communication forms the backbone of most meetings. The automatic transcription of speech in meetings is of crucial importance for meeting analysis, content analysis, summarisation, and analysis of dialogue structure. This paper presents the system for meeting transcription for the AMI project. The system was developed in a joint effort by the authors and others on the project and hence required very close international collaboration.

Widespread work on automatic transcription of speech in meetings started with yearly performance evaluations by the U.S. National Institute of Standards and Technology (NIST) with a first trial run in 2002. The Work was initially facilitated by the availability of the ICSI meeting corpus [9]. Further meeting resources were later made available by NIST [4] and Interactive System Labs (ISL) [1]. During NIST evaluations also recordings made at the Virgina Tech (VT) University were used. More recently two European projects, AMI and CHIL have collected and annotated substantial amounts of data. The AMI corpus [2] is now freely available.

1.1. Transcription of Meetings

Work on meeting transcription has has in part been dominated by the fact that the amount in-domain data is usually very small. As for any other spontaneous speech source, the cost of manual transcription is usually prohibitive. The number of speech resources for meetings is still small and most systems make use of adaptation of models from other domains. In [15] a recognition system for conversational telephone speech (CTS) formed the starting point, others have reported that bootstrapping from Broadcast News (BN) systems works well. In the following we also compare adaptation from CTS with unadapted training. in [8] we investigated whether meetings can be considered a domain, i.e. sufficiently uniform to warrant identical modelling of the language. We found that the vocabulary is very similar to that of BN with small out of vocabulary rates once BN is included. Meeting specific language models could give better perplexity and lower error rate, but with low margin.

Another issue is the recording source variability. Most corpora have audio recorded from individual head microphones (IHM), but ideally only microphones on the table, in microphone array configuration or not, should suffice for this task (multiple distant microphones, MDM). However, for MDM data a substantial performance degradation is observed.

The transcription system presented in this work makes use of the standard ASR framework, i.e hidden Markov model (HMM) based acoustic modelling and N-gram based language models (LMs). In the following we describe the main components of the AMI system for participation in the NIST RT06 evaluations. A brief description of the data used is followed by details on acoustic and language modelling. The final system architecture is described and results on the evaluation test set are presented.

2. DATA

In our initial work we found that including all above corpora into the training data helped [6]. However, the recording conditions differ considerably between corpora. The least difference is for IHM recordings where both lapel or head-mounted microphones of varying quality are used. This has an effect on the amount of noise and acoustic occlusion. For MDM microphones number and placement as well as quality differ. Whereas AMI and NIST have high quality microphone arrays, ICSI microphones where spread across the table, in approximately known location. For ISL and VT only one or two microphones were used.

For our IHM stem about 70 hours of speech data from the ICSI corpus [9], 13 hours from the NIST corpus [4] and 10 hours from ISL[1] were used. The AMI corpus collection was not completed at the time of system development and only 16 hours were included in the training set.The data consisted of data collected at IDIAP, and to a small extent at Edinburgh University [2]. This yielded a total training set (*ihmtrain05*) size of 108 hours.

For MDM the audio signal was enhanced (see Section 3.2) using selective bean-forming. Only a single speaker can be selected at each time. This approach cannot cope with overlapped speech and hence it had to be excluded from training. The final size of the training set (*mdmtrain05*) was 66 hours of speech.

3. PREPROCESSING

The audio pre-processing stages address several issues: The segmentation of the audio and implicit discarding of silence or noise;

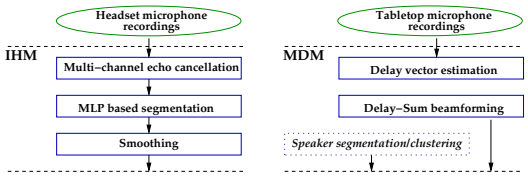

Fig. 1. Frontend Processing Stages for SAM and MDM.

the speaker labelling for later acoustic adaptation; the normalisation of input channels; and the suppression of noise. For IHM a system for segmentation and cross-talk suppression was developed. For MDM an enhancement based approach was taken were multiple channels were converted into a single channel consisting of the dominant speaker only. Segmentation and speaker labels were provided by ICSI/SRI [14]. Figure 1 shows the processing steps in diagramtic form. After the initial processing the audio signals are converted into feature streams, with vectors comprised of 12 MF-PLP features and raw log energy and first and second order derivatives are added. Cepstral mean and variance normalisation (CMN/CVN) is performed on a per channel basis.

3.1. Individual Head Microphones

Initially cross-talk suppression is performed using an adaptive LMS echo canceller followed by computation of 12 MF-PLP features. Additional features for the detection of cross-talk are extracted 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. The feature vectors are used to train a Multi-Layer-Perceptron (MLP) classifier with a 101 frame input layer, a 50 unit hidden layer and an output layer of two classes.the models are trained on 90 hours of data from all meetings in the *ihm-train05* set. On *rt05seval* the automatic segmentation gave equal performance as manual segmentation. More details can be found in [3].

3.2. Multiple Distant Microphones

Processing of MDM data takes account of the varying number of microphone channels and potentially unknown location of the microphones in relation to each other. The processing operates in several stages: First gain calibration is performed by normalising the maximum amplitude level of each of the input files. Then the background noise spectrum is estimated using the lowest energy frames in the recording and this is used to Wiener-filter the data to remove the stationary noise. In the next step delay vectors between channels are calculated on a per frame basis using generalised cross-correlation. Delays are computed in relation to a reference channel which also yields a relative scale factor. Delays and scale factors are then used in the final stage implementing superdirective beam-forming. More details can be found in [5].

While this approach is robust to a variety of configurations, for a small number of sparsely located microphones the estimates are very unreliable. In this case simply selecting the channel with the highest energy for every time frame was found to yield substantially lower word error rates (WERs).

4. ACOUSTIC MODELLING

All acoustic modells are based on cross-word state-clustered triphone models.It was found that, similar to CTS, 10-15% relative WER gain can be obtained using maximum likelihood based vocal tract length normalisation (VTLN) [6]. Secondly, heteroscedastic linear discrimant analysis (HLDA) gives consistent performance improvements [6]. Further gains could be obtained by discriminative training using the minimum phone error (MPE) criterion[11], also jointly with constrained maximum likelihood regression (MLLR) based speaker adaptive training (SAT). The left column of Table 1 shows WER results on *rt05seval*. In both cases substantial improvements are found.

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

Table 1. %WER results on *rt05seval* IHM (manual segmentation) with and without LCRC features.

4.1. Phoneme-State Posteriors Features

Recently there is increased interest in feature space representations that cover a long time span. Here we included features based on phone state posterior probability as computed by an MLP[13]. In order to generate the LCRC features 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 135D 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 25D feature vector is appended to the standard 39D feature vector.

Table 1 shows WER results both in comparison with standard features and with different training procedures. One can observe that the substantial gain from using these features is additive to other techniques. Similar patterns have been found on other test sets and MDM microphone data.

4.2. Adaptation to meetings

As shown in previous sections, word error rats on meeting data are still high. Part of the reason is the lack of sufficient training material. Hence adaptation of models trained on other domains is desirable and Maximum-A-Posteriori (MAP) based adaptation has been used successfully in this context (e.g. [14]). However, CTS operates on audio with reduced bandwidth. In [6] it was shown that better performance can be obtained using the full bandwidth available. As a consequence a MLLR based transform from narrow-band to wide-band data was derived and used in MAP adaptation of CTS models to meeting data. However, such a scheme is not viable with both HLDA and discriminative training. The solution to this problem was to project the meeting data into a narrowband space where both HLDA statistics can be gathered and discriminative training can 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 state clustering. The two sets of statistics are then combined in using MAP and 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 SAT and MPE training. Equivalently to adaptation of ML models with MAP, the JT/SAT/MPE models are adapted to meeting data using MPE-MAP[12]. The inclusion of SAT requires the presence of speaker transforms on meeting data. These are obtained from SAT training of MAP adapted CTS models in JT space. Table 2 shows results in JT space. A comparison of

Initial models	Adaptation	WER
CTS SAT MPE		30.4
CTS SAT MPE	ML-MAP	26.0
CTS SAT MPE + ML-MAP	MPE-MAP	23.9

Table 2. %WER results on *rt05seval* IHM with adaptation from CTS to *ihmtrain05*.

	TOT	AMI	CMU	ICSI	NIST	VT
baseline	53.6	46.5	50.2	48.2	53.6	63.0
all channel	54.7	48.4	51.3	49.4	55.1	63.3
CHAT	52.9	47.2	48.0	47.1	52.0	63.3

Table 3. %WER results on *rt05seval* MDM with automatic segmentation.

the final WER results with that in Table 1 shows a 0.6% absolute improvement. However, the elaborate process prohibited inclusion of LCRC features at this point. A more detailed analysis of this procedure can be found in [10].

4.3. Channel Adaptive Training Experiments

For MDM training the amount of training data is even smaller, due to the constraint of using only non-overlapped speech. Furthermore speech enhancement introduces additional distortions that are not modelled appropriately. One approach to address this issue is by training on all microphone channels (as used in [14]). Table 3 shows that performance degraded. However, when using a SAT style training on each microphone channel (CHAT), i.e. one set of CMLLR transforms per channel, a small performance gain was observed. Table 4 shows that with VTLN, HLDA, and discriminative training a moderate gain is retained. Note that decoding is till performed on the enhanced single audio channel.

5. LANGUAGE MODELLING AND VOCABULARY

The UNISYN pronunciation lexicon forms the basis of dictionary development with pronunciations mapped to the General American accent [6]. The original phoneme set was mapped to 45 phonemes which introduced some unusual pronunciations for words including flaps. However experiments did not suggest that this was problematic. Normalisation of lexicon entries to resolve differences between American and British derived spelling conventions was performed. Pronunciations for a further 15000 words were generated manually for work in this paper.

Standard N-gram language models (LMs) were built using the SRI LM toolkit [1]. Table 5 lists the text resources used for training of bigram, trigram and 4-gram LMs. Note that in contrast to other web-data, the AMI web-data was collected using techniques to collect text that is different to the already existing background material [16]. From the interpolation weights it is clear that conversational data is most important. The perplexity of the interpolated was 84.3 for the interpolated trigram and 81.2 for the 4-gram model on *rt06seval*.

6. SYSTEM OVERVIEW AND PERFORMANCE

The AMI 2006 system operates in a total of six passes (see Figure 2). 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 on the description of the IHM system and highlight the differences for MDM later on.

In the first pass, P1, the front-end converts the recordings into feature streams as described in Section 3. The audio stream is split into meaningful segments. 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

[1] `http://www.speech.sri.com/projects/srilm`

	ML	MPE
enhancement-based	42.9	40.0
CHAT	43.0	s 39.0

Table 4. %WER results on *rt05seval* with reference segmentation, VTLN and HLDA.

LM component	size	weights (trigram)
AMI data from rt05s	206K	0.038
Fisher	21M	0.237
Hub4 LM96	151M	0.044
ICSI meeting corpus	0.9M	0.080
ISL meeting corpus	119K	0.091
NIST meeting corpus	157K	0.065
Switchboard/callhome	3.4M	0.070
webdata (meetings)	128M	0.163
webdata (fisher)	128M	0.103
webdata (AMI)	138M	0.108

Table 5. Language model data set sizes and weights in interpolation.

that are subsequently used for estimation of VTLN warp factors. The acoustic models *M1* are ML models trained on *ihmtrain05* only. An interpolated trigram LM is used in decoding. The feature vectors and CMN and CVN are subsequently recomputed and the LCRC features are appended.

The *M2* models are trained on *ihmtrain05* using VTLN, HLDA, SAT and MPE and LCRC features. They are adapted using the transcripts of the first pass and a single CMLLR transform. Here the interpolated bigram LM is used to to output the bigram word lattices in the second pass, P2. In the third pass (P3) the lattices are expanded using the 4-gram LM. These lattices are subsequently used for acoustic rescoring in all passes.

The third model set, *M3*, is identical to the one described in 4.2. In the fourth pass both *M2* and *M3* models are adapted using both CMLLR and MLLR with regression class trees for up to four classes. Lattices are rescoring using the adapted models and a dictionary containing pronunciation probabilities (PPROB). The output of P4a is used for adaptation of M3 models and the output of P4b is used to adapt the M2 models in the same fashion. Finally confusion networks (CNs) are generated, combined and the sequence with the highest local posterior probability is extracted[7]. During development only unreliable improvement was found by combination and this stage was hence omitted.

Table 6 shows results for each processing stage. A large difference in WER can be observed by stepping from P1 to P2. After the third pass the results are already very close to the final performance. Even though the P4b system has lower performance on its own the inclusion into the adaptation path yields a further 0.5% WER absolute. Simple adaptation with P4a supervision did not give any improvement. Also, note the similarity in WERs across all meeting corpora. The difference between IHM and MDM lies in the front-end, the acoustic model training set, and that only *M1* and *M2* acoustic models are used. The MDM performance is given in Table 7 (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 there is a considerable difference between performance on IHM and MDM.

7. CONCLUSIONS

We have presented the AMI 2006 system for transcription of speech in meetings. Compared to our initial system a WER reduction on IHM of more than 20% relative and competitive performance in the NIST RT'06 evaluations was achieved. The MDM performance still is comparatively poor and requires better modelling. One of the main

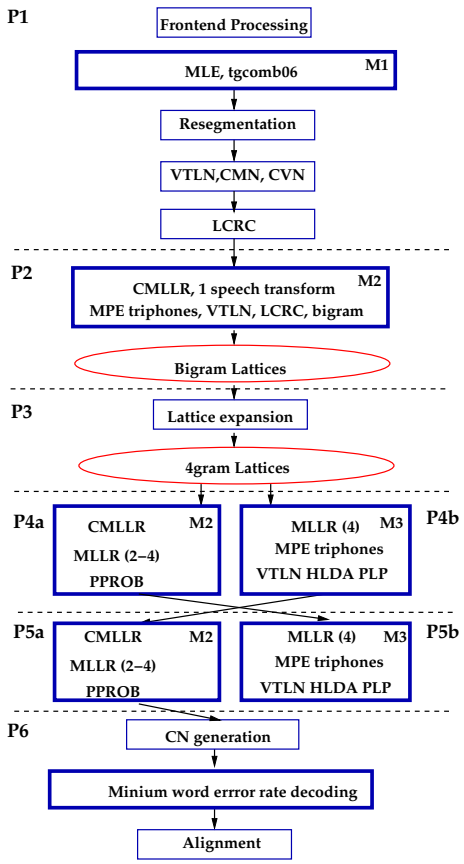

Fig. 2. Processing stages of the complete AMI 2006 system.

issues is poor use of training data. Schemes such as CHAT can be used to improve utilisation of the available data and robust adaptation from other domains will allow better smoothing. In combination this should substantially lower error rates.

8. 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.

9. REFERENCES

[1] S. Burger, V. MacLaren, and H. Yu. The ISL meeting corpus: The impact of meeting type on speech style. In *Proc. ICSLP*, 2002.

[2] 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. The AMI meeting corpus. In *Proc. MLMI'05*, Edinburgh, 2005.

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

[4] J. Garofolo, C. Laprun, M. Miche, V. Stanford, and E. Tabassi. The nist meeting room pilot corpus. In *Proc. 4th Intl. Conf. on Language Resources and Evaluation*, 2004.

[5] 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 *Proc. NIST RT'05 Workshop*, Edinburgh, 2005.

	TOT	CMU	EDI	NIST	TNO	VT
P1	42.0	41.9	41.0	39.0	42.1	44.8
P2a	29.2	29.2	27.4	27.7	29.5	32.4
P3	26.0	25.7	24.6	25.2	26.3	29.5
P4a	25.1	25.0	22.8	23.8	26.0	29.1
P4b	25.6	25.3	23.8	24.9	24.3	29.8
P5a	24.6	24.4	22.6	23.6	24.1	28.8
P5a-cn	24.2	24.0	22.2	23.2	23.6	28.2

Table 6. %WER results of the AMI 2006 system on *rt06seval* IHM. EDI denotes AMI data collected at Edinburgh University, TNO denotes AMI data collected at TNO.

	TOT	Sub	Del	Ins
P1	58.2	35.8	16.7	5.7
P2a	45.6	26.4	15.1	4.1
P3	42.0	24.5	13.2	4.4
P4a	41.7	22.9	14.9	3.9
P5	40.9	22.2	15.3	3.5

Table 7. %WER results on *rt06seval* MDM.

[6] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals. The development of the AMI system for the transcription of speech in meetings. In *Proc. MLMI'05*, 2005.

[7] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan. The ami meeting transcription system : Progress and performance. In *Proc. NIST RT'06 Workshop*, Springer LNCS, 2006.

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

[9] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The ICSI meeting corpus. In *Proceedings IEEE ICASSP*, 2003.

[10] M. Karafiat, L. Burget, J. Cernocky, and T. Hain. Application of cmllr in nb-wb adapted system. In *Submission ICASSP*, 2007.

[11] D. Povey. *Discriminative Training for Large Vocabulary Speech, Recognition*. PhD thesis, Cambridge University, July 2004.

[12] D. Povey, M. J. F. Gales, D. Y. Kim, and P. C. Woodland. MMI-MAP and MPE-MAP for acoustic model adaptation. In *Proc. Eurospeech'03*, 2003.

[13] P. Schwarz, P. Matjka, and J. Cernock. 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, page 8, Brno, 2004.

[14] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, F. Grezl, A. Janin, A. Manda, B. Peskin, C. Wooters, and J. Zheng. Further progress in meeting recognition: The ICSI-SRI Spring 2005 Speech-to-Text evaluation system. In *Proc. NIST RT'05 Workshop.*, 2005.

[15] A. Stolcke, C. Wooters, N. Mirghafori, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin, and M. Ostendorf. Progress in meeting recognition: The ICSI-SRI-UW spring 2004 evaluation system. In *Proc. NIST RT04S Workshop.*, 2004.

[16] V. Wan and T. Hain. Strategies for language model web-data collection. In *Proc. ICASSP'06*, number SLP-P17.11, 2006.