The Development of the AMI System for the Transcription of Speech in Meetings

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

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

Abstract. The automatic processing of speech collected in conference style meetings has attracted considerable interest with several large scale projects devoted to this area. This paper describes the development of a baseline automatic speech transcription system for meetings in the context of the AMI (Augmented Multiparty Interaction) project. We present several techniques important to processing of this data and show the performance in terms of word error rates (WERs). An important aspect of transcription of this data is the necessary flexibility in terms of audio pre-processing. Real world systems have to deal with flexible input, for example by using microphone arrays or randomly placed microphones in a room. Automatic segmentation and microphone array processing techniques are described and the effect on WERs is discussed. The system and its components presented in this paper yield competive performance and form a baseline for future research in this domain.

1 Introduction

Many people spend a considerable time in their working life in meetings, however the efficiency of meetings is often low and hence approaches for streamlining the process and for retaining and crystallising the right information have been developed. So far computers are rarely used to aid this process. Projects like AMI (which stands for 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, information is presented for example in the form of presentation slides, drawings on boards, and of course by verbal communication. The latter 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. Widespread Work on automatic recognition of speech in meetings started with yearly performance evaluations by the U.S. National Institute of Standards and Technology (NIST) [19]. Work on meeting transcription was initially facilitated by the collection of the ICSI meeting corpus [13] which was followed by trail NIST meeting transcription evaluations in Spring 2002. Further meeting resources were made available by NIST [9], Interactive System Labs (ISL) [3] and the Linguistic Data Consortium RT04s Meeting evaluations [19].

As the number of speech resources for meetings is still relatively small, similar to work presented in [22], a recognition system for conversational telephone speech (CTS) forms the starting point for our work on meetings. This approach was preferred to bootstrapping from Broadcast News (BN) systems (as for example in [21]) as the meeting style is expected to be colloquial rather than presentational. In the following we give a description of meeting resources followed by a description of our CTS baseline system. This is followed by an analysis of meeting vocabulary and linguistic context followed by experimental results with various approaches to acoustic modelling.

2 Meeting resources

The ICSI Meeting corpus [13] is the largest meeting resource available consisting of 70 technical meetings at ICSI with a total of 73 hours of speech. The number of participants is variable and data is recorded from head-mounted and a total of four table-top microphones. A 3.5 hour subset of this corpus covering 30 minute extracts of 7 meetings was set aside for testing (icsidev). Further meeting corpora were collected by NIST [9] and ISL [3], with 13 and 10 hours respectively.Both NIST and ISL meetings have free content (e.g. people playing games or discussing sales issues) and number of participants. We make also make use of the RT04s NIST evaluation set (rt04seval) which also includes meetings recorded by the LDC.

As part of the AMI project a major collection and annotation effort of the AMI meeting corpus[4] is currently underway. Data is collected from three different model meeting rooms in Europe (mostly Edinburgh and IDAP at the moment). Overall more than 100 hours of speech are to be transcribed. The meeting language is English. Each meeting normally has four participants and the corpus will be split into a *scenario* portion and individual meetings. The scenario portion will involve the same participants over multiple meetings on one specific task. The data used in this paper only originates from scenario meetings. An additional development set (amidev) consisting of 8 meetings from 2 locations is used for testing.

3 The AMI CTS system

All systems in this paper are based on standard speech recognition technology such as HMM based acoustic models and N-gram based language models. In the following we briefly outline the front-end and acoustic modelling, dictionary consruction, and language modelling on this task.

3.1 Acoustic modelling

Font-ends make use of 12 MF-PLP [24, 12] coefficients and the 0th cepstral coefficient c_0. These are derived from a reduced bandwidth of 125-3800Hz. First and second order derivatives are added to form a 39 dimensional feature vector. Cepstral mean and variance normalisation is performed on complete conversation sides and hence are implicitly speaker specific. Acoustic models are phonetic decision tree state clustered triphone models with standard left-to-right 3-state topology. They were obtained using standard HTKmaximum likelihood training procedures (see for example [11]). The system uses approximately 7000 states where each state is represented as a mixture of 16 Gaussians. Speaker adaptive training is performed in the form of vocal tract length normalisation (VTLN) both in training and test. Warp factors are estimated using a parabolic search procedure, a piecewise linear warping function and a maximum likelihood criterion[11]. Speaker adaptation is perfermed using maximum likelihood linear regression (MLLR) of the means and variances[8].

Feature transformation is applied in the form of smoothed heteroscedastic linear discriminant analysis (SHLDA) [17]. SHLDA is used to reduce a 52 dimensional formed by the standard feature vector plus third derivatives to 39 dimensions. HLDA estimation procedure[16] requires the estimation of full covariance matrices per Gaussion. SHLDA uses smoothing of the covariance estimates by interpolating with standard LDA type with-in class covariances.

$$\boldsymbol{\Sigma}_{sm} = \alpha \boldsymbol{\Sigma} + (1 - \alpha) \boldsymbol{\Sigma}_{WC} \tag{1}$$

$\boldsymbol{\Sigma}_{sm}$ is the smoothed estimate of the covariance matrix and $\boldsymbol{\Sigma}_{WC}$ is the LDA type within-class matrix estimate based on an occupancy weighted average. Values for α of $0.8 - 0.9$ were found to yield satisfactory results.

3.2 Dictionaries

The UNISYN pronunciation lexicon [7] forms the basis of dictionary development with pronunciations mapped to the General American accent. Normalisation of lexicon entries to resolve differences between American and British derived spelling conventions was performed yielding a 115k word base dictionary. Pronunciations for a further 11500 words were generated manually for work in this paper. For consistency and a simplified manual pronunciation generation process hypotheses generation procedures have been developed. Pronunciations for partial words are automatically derived from the baseform dictionary. Hypotheses for standard words were generated using CART based letter-to-sound rules. The CART based letter-to-sound prediction module was trained on the UNISYN dictionary using tools provided with the Festival speech synthesis software [1] using left and right context of five letters and left context of two phones. This gave 98% phone error rate and 89% word error rate on the base dictionary., for

Table 1. Size of various text corpora in million words (MW).

Corpus name	#words (MW)
Swbd/CHE	3.5
Fisher	10.5
Web (Swtchboard)	163
Web (Fisher)	484
Web (Fisher topics)	156
BBC - THISL	33
HUB4-LM96	152
SDR99-Newswire	39
Enron email	152
ICSI meeting	1
Web (meetings)	128

Table 2. Perplexities on the NIST Hub5E 1998/2001/2002 evaluation test sets.

Hub5e eval sets	Bigram	Trigram	4-gram
Swbd	104.53	85.97	84.12
Swbd + HUB4	95.00	72.55	69.04
Swbd + HUB4 + Web	90.89	66.75	61.59

manually generated pronunciations the error rates were 89% and 51% respectively. Although the word accuracy is quite low on new words (many of which were proper names, partial words etc.), the phone accuracy remains relatively high.

3.3 Language modelling and Vocabulary

Selection of vocabulary for recognition is based on a collection of in-domain words. However, in the case of insufficient data it is beneficial to augment this list with the most frequent words from other sources, for example Broadcast News (BN) corpora. This "padding" technique was used for all dictionaries in this paper unless stated otherwise. The target dictionary size was 50000 words and the source of words was BBC news data, the Broadcast News 1996 Hub4 corpus (HUB4-LM96), and Enron data[14] (see table 2).

Language model training data for conversational speech is sparse. Hence models are constructed from other sources and interpolated (as in e.g. [11]). This is true for both CTS and meeting data. Hence we have processed a large number of different corpora to form the basis of our language models. The most important corpora are listed in Table 1. A full discussion of all source material would go beyond the scope of this paper. The most important non-standard data was found to be the the Web collected resources [2] and ICSI meetings. In total more than 1300 MW of text are used. Each corpus was normalised using identical processes. Apart from standard cleanup we tried to ensure normalised spelling and uniform hyphenations across all corpora. For the training and testing of language models the SRI LM toolkit [23] was used to train models with Kneser-Ney discounting and Backoff. Table 2 shows perplexity results on the NIST Hub5e evaluation sets. Note the substantial reduction in perplexity by the additional web resources.

Table 3. %WER results on the NIST Hub5E 2001 evalution set.

eval01	VTLN	MLLR	non-HLDA	SHLDA
pass1			37.2	35.0
pass2	×		33.8	32.1
pass3	×	×	32.1	30.6

Table 4. Statistics for meeting corpora.

	ICSI	NIST	ISL	AMI
Avg. Dur (sec)	2.42	3.98	3.21	3.95
#words	823951	157858	119184	154249
#unique wds	11439	6653	5622	4801

3.4 Decoding and overall system performance

Decoding operates in three passes. The Cambridge University speech decoder HDecode is used for recognition with trigram language models. Table 3 shows results for each pass. The first pass yields a first level transcription which is used or VTLN warp factor estimation. In the second pass improved output is generated using VTLN trained models. The final output is obtained after MLLR adaptation using transforms for speech and silence. The table also gives a comparison of results with and without SHLDA. Trigram language models as described above were used in the experiments. A significant reduction in word error rate (WER) from both VTLN and SHLDA is observed.

4 Language in Meetings

Even though of general conversational nature, meeting data differs substantially from CTS. First of all the acoustic recoding condition is usually more complex as the speaker has no feedback on the recording quality. Speech signals of close-talking microphones are distorted by heavy breathing, head-turning and cross-talk. Table 4 shows raw statistics on several meeting corpora. Average utterance durations are larger than on CTS, however with great variation. We can also observe that corpus size is not a good predictor for the number of unique words in the corpus and hence complexity.

4.1 Vocabulary

We shall loosely define a domain as a set of sub-corpora that, when used in a combined non-discriminative fashion, yield better performing models than the parts. This definition is not strict and will show a tendency to combine small corpora. However for the purpose of model training the question of how to use data is most important. Table 5 shows on the left hand side Out Of Vocabulary (OOV) rates using vocabulary derived from each meeting corpus. The OOV rates do not correlate perfectly with vocabulary sizes (Table 4). On the right hand side the wordlists are padded as described in section 3.3. It is evident that overall the effect of vocabulary mismatch is greatly reduced uniformly for all cases. This suggest that only a very small amount of meeting specific vocabulary is necessary. Hence padding was used in all further experiments.

Table 5. %OOV rates of meeting resource specific vocabularies. Columns denote the word list source, rows the test domain.

	No padding				Padding to 50k			
	ICSI	NIST	ISL	AMI	ICSI	NIST	ISL	AMI
ICSI	0.00	4.95	7.11	6.83	0.01	0.47	0.58	0.57
NIST	4.50	0.00	6.50	6.88	0.43	0.09	0.59	0.66
ISL	5.12	5.92	0.00	6.68	0.41	0.37	0.03	0.57
AMI	4.47	4.39	5.41	0.00	0.53	0.53	0.58	0.30
COMBINED	1.60	4.35	6.15	5.98	0.16	0.42	0.53	0.55

Table 6. Cross meeting room perplexities on subsets of of rt04seval and rt05samidev. COMBINED denotes training or testing using all meeting data.

Test Corpus	ICSI	NIST	ISL	AMI	COMBINED
ICSI	68.17	74.57	73.76	77.14	67.97
NIST	105.91	100.87	102.01	105.95	101.25
iSL	104.68	99.45	98.45	106.39	102.86
AMI	115.56	114.26	114.41	88.91	94.08
LDC	97.78	90.66	88.87	92.44	93.84
COMBINED	107.46	105.93	105.73	90.62	92.74

4.2 Content

Apart from the raw word difference it is important understand the effect of the wide range of topics covered in the various meetings. A set of experiments was conducted to compare meeting resource optimised language models on the basis of the meeting resource specific (MRS) padded vocabularies. Language models are obtained by optimisation of interpolation weights for the components outlined in Table 1. Table 6 shows perplexities on all corpora. In all cases that the best perplexities are achieved on the originating corpus, however with little margin. Note also that the MRS LMs significantly outperform the generic LMs only in the case of ISL and AMI. In general the perplexity of ICSI test data is very low. This appears to be a property of this data set.

5 Meeting transcription

Common for all meeting rooms is that audio is recorded either by close-talking microphones or via single or multiple distant microphones. The latter may be arranged in a fixed array configuration. Due to interaction between speakers the system must be capable of speech detection and and speaker grouping as well as recognition. In the following we first outline techniques for audio segmentation and microphone array processing, followed by a description of model training procedures and recognition results.

5.1 Automatic segmentation

Speech activity detection (SAD) for close talking microphones poses a significant challenge. The high levels of cross-talk and non-speech noise (such as breath or contact noise) prohibit the use of threshold based techniques, the standard in more ‘friendly’ recording conditions. The system used here is a straight-forward statistical based approach with additional components to control cross-talk between channels. Statistical approaches to SAD typically use HMM or GMM based classifiers with special feature vectors such as channel cross-correlation and kurtosis (e.g. [20, 25]). A 14 dimensional PLP [12] feature vector is used to train a Multi-Layer-Perceptron (MLP) classifier with a 101 frame input layer, a 20 unit hidden layer and an output layer of two classes. Parameters are trained on 10 meetings from each meeting resource totalling around 20 hrs of data. Further 5 meetings from each corpus are used to determine early stopping of the parameter learning. The utterance segmentation uses Viterbi decoding and scaled likelihoods derived from the MLP and a minimum segment duration of 0.5 seconds.

Cross talk suppression is performed at the signal level using adaptive-LMS echo cancellation[18]¿ Additons to the basic system are: the use of multiple reference channels in cancellation; automatic channel delay estimation and offsetting of reference signals to account for this delay; automatic cross-talk level estimation; and ignoring of channels which produce low levels of cross-talk. Updates are further made on a per sample basis to account for non-stationary ‘echo’ path. On the classifier level additional features were introduced to aid the detection of cross-talk:

$$\mathrm{RMS}_{norm}\left(x_{t-L}^{t+L}(i)\right) = \log\left(\mathrm{RMS}\left(x_{t-L}^{t+L}(i)\right)\right) - \log\left(\sum_{j=1}^{N} \mathrm{RMS}\left(x_{t-L}^{t+L}(j)\right)\right), \quad (2)$$

$$Kur\left(x_{t-L}^{t+L}\right) = \frac{E\left\{\left(x_{t-L}^{t+L} - E\left\{x_{t-L}^{t+L}\right\}\right)^{4}\right\}}{E\left\{\left(x_{t-L}^{t+L} - E\left\{x_{t-L}^{t+L}\right\}\right)^{2}\right\}^{2}}, \quad (3)$$

$$Cep\left(x_{t-L}^{t+L}\right) = \max_{t=P_l-P_h}\left(\mathcal{F}\left(\log\left(\left|\ \mathcal{F}\left(x_{t-L}^{t+L}\right)\right|\ \right)\right)\right). \quad (4)$$

where x_{t-L}^{t+L} is the signal x windowed over $2 \cdot L$ samples and P_l and P_h are the minimum and maximum pitch period over which peak picking is carried out (corresponding to 50-300Hz). Eq. 2 describes across-meeting normalised RMS energy, Eq. 3 signal and spectrum kurtosis, and Eq. 4 as a voicing strength measure based on the maximum amplitude in the speech cepstrum in the range of frequencies 50-300Hz.

5.2 Microphone array processing

Audio from multiple distant microphones (MDMs) can be used in variety of ways. The AMI baseline system uses an enhancement based approach. Recordings from a number of microphones placed in the meeting rooms are combined to arrive at a single, enhanced output file that is then used as input for recognition. The system is required to cope with a number of unknown variables: varying numbers of microphones; unknown microphone placement; unknown numbers of talkers; time variant skew between input channels introduced by the recording system; and different room geometry and acoustic conditions.

The MDM processing operates in a total of four stages. First gain calibration is performed by normalising the maximum amplitude level of each of the input files. Then a noise estimation and removal procedure is run. This in itself is a two pass process. On the first pass the noise spectrum $\Phi_{nn}(f)$ of each input channel is estimated as the noise power spectrum of the M lowest energy frames in the file ($M = 20$ was used for the current experiments). On the second pass a Wiener filter with transfer function $\frac{\Phi_{xx}(f)-\Phi_{nn}(f)}{\Phi_{xx}(f)}$ (where $\phi_{xx}(f)$ is the input signal spectrum) is applied to each channel to remove stationary noise. The noise coherence matrix Q, estimated over the M lowest energy frames, is also output at this time. In the third stage delay vectors between each channel pair are calculated for every frame in the input sample. The delay between two channels is the time difference between the arrival of the dominant sound source and is calculated by finding the peak in the Generalised Cross Correlation [15] between input frames across two channels. The delay vector is given as the delays for all pairs with respect to a single reference channel - there are therefore N delays in each vector, with the delay for the reference channel equal to 0. Further a vector of relative scaling factors is calculated, corresponding to the ratio of of frame energies between each channel and the reference channel. The start and end times in seconds, along with the delay and scaling factors are output for each frame. Finally The delay and scaling vectors are then used to calculate beamforming filters for each frame using the standard superdirective technique [5, 6]. The superdirective formulation requires knowledge of the noise coherence matrix. However this is not available as the microphone positions are not known. Either a unity coherence matrix may be used (leading to delay-sum filters) or the Q matrix estimate in the second stage may be used. Each frame is then beamformed using the appropriate filters and the output subsequently used for recognition.

5.3 Model building

As outlined above the the fact that meeting resources are still comparatively small, bootstrapping from CTS models was used. However, as CTS data is only available at a bandwidth of 4kHz this poses additional questions on the initialisation and training procedure.

Bandwidth and Adaptation Table 7 shows recognition performance on the icsidev test set using various model training strategies. The baseline CTS systems yield a still reasonable error rate. Training on 8kHz-limited (NB) ICSI training data yields a WER of 27.1%. Using the full bandwidth (WB) reduces the WER by 1.8%. The standard approach for adaptation to large amounts of data is MAP [10]. As CTS is NB only, adaptation to WB ICSI data was performed using MAP adaptation in an iterative fashion. However the performance of the adapted NB system was still poorer than that of the system trained on WB data. The results show that MAP adaptation from CTS models while using wideband data is desirable. In our implementation the adaptation model set is used for two purposes: for computation of state level posteriors and to serve as a prior.

Table 7. %WER results on icsidev for several different training strategies and a trigram LM optimised for the ICSI corpus.

data	bandwidth	adapt	#iter	%WER
CTS	NB	-	-	33.3
ICSI	NB	-	-	27.1
ICSI	WB	-	-	25.3
ICSI	NB	MAP	1	26.5
ICSI	NB	MAP	8	25.8
ICSI	WB	MLLR + MAP	8	24.6
ALL	WB	MLLR + MAP	8	25.8

Table 8. %WER on the rt04eval sets . TOT gives WERS overall, while MRS denotes the use of language models focusing on specific meeting rooms

	TOT	ISL	ICSI	NIST	LDC
MRS ISL	40.2	44.7	25.8	34.1	53.8
MRS ICSI	40.2	45.2	25.1	34.7	53.5
MRS NIST	40.2	44.6	26.2	34.1	53.6
MRS AMI	41.0	45.1	26.9	35.8	54.2
COMBINED	40.0	44.5	25.6	34.4	53.4

Even if the former is performed well, NB models cannot be used to serve as prior directly. In order to overcome this problem the means of the CTS models were modified using block-diagonal MLLR transforms. One transform for speech and one for silence was estimated on the complete ICSI corpus using models trained on ICSI NB data. After an initial step with MLLR-adapted CTS models iterative MAP adaptation is resumed as before. The use of more detailed modelling of the transition from NB to WB by the use of more transforms was not found to yield a significant performance improvement. After 8 iterations a further 0.9% reduction in WER is obtained.

Meeting resource specific language modelling The language and vocabulary in meetings differs substantially. We have found evidence that his is also true for the acoustics However the advantage of having more data outweighs the differences. Hence we use acoustic models trained on the all meeting resources. Table 8 shows WER results using acoustic models trained on the complete meeting data and specific language models. An initial observation makes clear that on average the best strategy is to combine all the resources (similar to the acoustics). Further the variation of scores is modest whereby AMI data is distinct from all other resources. A moderate beneficial effect can be observed from using meeting room specific language models.

IHM Processing The sections above gave an outline of the components required for a baseline system on meeting transcription. The task of combining the components in a sensible complex. For optimal performance many of the techniques cannot just simply be "plugged" together. Table 9 shows WER results using various model building techniques. Models are trained on a total of 96 hours of meeting speech. The baseline model yields 40% overall. By far

Table 9. %WER on the rt04eval set using a combined tigram language model. CTS denotes CTS-adapted, EC echo cancellation.The table shows gender specific results (F/M) and results per meeting room . In the first section the reference segmentation of the data is used.

Name	CTS	VTLN	EC	TOT	F	M	ISL	ICSI	LDC	NIST
BASE	×			40.0	39.4	40.4	44.5	25.6	53.4	34.4
VTLN1	×	×		36.9	36.4	37.2	42.0	22.4	50.3	30.5
VTLN2		×		37.6	36.0	38.4	42.7	23.3	51.3	30.1
VTLN1 - SHLDA	×	×		36.0	35.1	36.5	41.0	21.8	50.5	27.4
EC1	×		×	40.3	39.5	40.7	44.7	25.9	54.8	33.1
VTLN-EC1	×	×	×	37.0	36.1	37.5	41.2	22.9	50.8	30.9
SEG1	×			50.8	51.1	50.6	50.4	38.2	73.3	37.4

Table 10. %WER on rt04seval and rt05samidev-n when training on various meeting resource combinations.

	rt05seval					rt05samidev-n		
	TOT	ISL	ICSI	LDC	NIST	TOT	UEDIN	IDIAP
. ICS,NIST	50.4	56.2	24.1	61.1	36.9	59.1	60.2	58.4
ICSI,NIST,ISL	50.6	56.2	22.9	61.8	37.2	59.1	60.0	57.6
ICSI,NIST,ISL,AMI	50.3	54.5	27.4	61.3	36.2	57.3	59.0	54.5

the best performance is achieved on the ICSI portion of the data and performance is roughly gender balanced. Similar to CTS the use of VTLN yields a substantial improvement. Comparing the systems VTLN1 and VTLN2, the gain from CTS-adaptation remains even in conjunction with VTLN. The next part of the table shows the use of echo-cancelled (EC) data (as used for segmentation). Virtually no effect on recognition performance can be observed. The last section shows results with automatic segmentation (all other results are based on reference segmentation). The first system, SEG1, only makes use of the basic configuration, i.e. using an MLP only trained on PLP features.

MDM processing Almost all meeting corpora used a different approach to record speech with remote microphones. In the ICSI corpus microphones are not in fixed array configuration, the ISL corpus only uses one distant microphone, AMI uses a circular microphone array. Table 10 shows performance results with models trained on specific corpora. Overall the size and type of data used appears to have little impact on performance. Only the use of AMI training data appears to aid recognition on the AMI test set. The enhancement based approach described in section 5.2 has the disadvantage that it cannot cope with overlapped speech. Since straight-forward removal of overlapping segments however would be far to restrictive. Instead word timings from forced alignment were used to identify overlaps. Speech segments were split, either at point of at least 100ms silence (ms10), of silence occurrence(ms0), or at arbitrary word boundaries (wb). These approaches reduce the original training set size of 96 hours to 56, 63 or 66 hours respectively. Table 11 shows associated WER results. Only a minor preference of an increase in training set size is evident. However training set size has an impact on the effect of channel based normalisation schemes. 11 shows

Table 11. %WER on rt04seval and rt05samidev-n with different amounts of traiing data. ms0, ms10,and wb describe data preparation (see text).

	rt05seval					rt05samidev-n		
	TOT	CMU	ICSI	LDC	NIST	TOT	UEDIN	IDIAP
ms0	51.0	55.4	26.4	63.4	34.9	57.4	58.9	55.0
ms10	51.0	54.3	25.9	63.6	37.0	56.4	58.0	54.0
wb	50.7	56.5	24.3	61.9	36.4	56.3	58.2	53.4
VTLN - wb	47.2	51.4	20.6	60.2	31.3	-	-	-
wb icsiseg	55.2	59.5	32.2	66.7	40.5	-	-	-

the performance after VTLN in both training and test, yielding improvements comparable to IHM. Finally table 11 shows results for use of automatic segments as generated by the ICSI segmenter[22] which results in 5% absolute reduction in WER, mostly driven by an increase in the deletion rate. note that the greatest degradation was on the ICSI corpus.

6 Conclusions

In this paper the components of the AMI meeting transcription system were described. So far the system is equipped with baseline compomen ts that allow the processing of the highly variable data. We have shown: the feasibility to use the Edinburgh UNISYN dictionary for speech recognition, the effective use of language model data for meetings collected from the internet; the effective use of SHLDA and VTLN on CTS and meetings, both in IHM and MDM recorddings; the language properties of meeting rooms; and effective data preparation for this domain. We have further presented initial transcription results on the AMI meeting corpus.

7 Acknowledgements

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

References

1. A.W. Black, P. Taylor and R. Caley (2004). The Festival Speech Synthesis System, Version 1.95beta. CSTR, University of Edinburgh, Edinburgh.
2. I. Bulyko, M. Ostendorf and A. Stolcke. Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures. in Proc HLT'03.
3. S. Burger, V. MacLaren, H. Yu (2002). The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style. In Proc. ICSLP'2002.

4. 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, D. Reidsma (2005. The AMI Meeting Corpus. Submitted to MLMI'05.
5. H. Cox, R. Zeskind, and I. Kooij (1986). Practical supergain. IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-34(3):393–397.
6. H. Cox, R. Zeskind, and M. Owen (1987). Robust adaptive beamforming. IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-35(10):1365–1376.
7. S. Fitt (2000). Documentation and user guide to UNISYN lexicon and post-lexical rules, Tech. Rep., Centre for Speech Technology Research, Edinburgh.
8. M.J.F. Gales & P.C. Woodland (1996). Mean and Variance Adaptation within the MLLR Framework. *Computer Speech & Language*, Vol. 10, pp. 249–264.
9. J.S. Garafolo, C.D. Laprun, M. Michel, V.M. Stanford, E. Tabassi (2004). In Proc. 4th Intl. Conf. on Language Resources and Evaluation (LREC'04).
10. J.L. Gauvain, C. Lee (1994). MAP estimation for multivariate Gaussian mixture observation of Markov Chains, IEEE Tr. Speech& Audio Processing, 2, pp. 291-298.
11. T. Hain, P. Woodland, T. Niesler, and E. Whittaker (1999). The 1998 HTK system for transcription of conversational telephone speech. Proc. IEEE ICASSP, 1999.
12. H. Hermansky (1990). Perceptual Linear Predictive (PLP) analysis of speech. Acoustical Society of America, 87(4):1738–1752.
13. A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters (2003). The ICSI Meeting Corpus. ICASSP'03, Hong Kong.
14. B. Klimt, Y. Yang (2004). Introducing the Enron Corpus, Second Conference on Email and Anti-Spam, CEAS 2004.
15. C. H. Knapp and G. C. Carter (1976). The generalized correlation method for estimation of time delay/ IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-24:320–327, August 1976.
16. N. Kumar (1997), Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition.PhD thesis, John Hopkins University, Baltimore.
17. L. Burget (2004), Combination of Speech Features Using Smoothed Heteroscedastic Linear Discriminant Analysis. in Proc. ICSLP'04, Jeju island, KR, 2004, p. 4.
18. D. Messerschmitt, D. Hedberg, C. Cole, A. Haoui, and P. Winship (1989). Digital voice echo canceller with a TMS32020. Appl. Rep. SPRA129, Texas Instruments.
19. Spring 2004 (RT04S) Rich Transcription Meeting Recognition Evaluation Plan. NIST, US. Available at `http://www.nist.gov/speech`.
20. T. Pfau and D.P. W. Ellis (2001). Hidden markov model based speech activity detection for the ICSI meeting project. Eurospeech'01.
21. T. Schultz, A. Waibel, M. Bett, F. Metze, Y. Pan, K. Ries, T. Schaaf, H. Soltau, M. Westphal, H. Yu, and K. Zechner (2001). The ISL Meeting Room System. In Proc. of the Workshop on Hands-Free Speech Communication (HSC-2001), Kyoto.
22. A. Stolcke, C. Wooters, N. Mirghafori, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin, and M. Ostendorf (2004). Progress in Meeting Recognition: The ICSI-SRI-UW Spring 2004 Evaluation System. NIST RT04 Workshop.
23. The SRI Language Modelling Toolkit (SRILM). `http://www.speech.sri.com/projects/srilm`, SRI international, California.
24. P.C. Woodland, M.J.F. Gales, D. Pye & S.J. Young (1997). Broadcast News Transcription using HTK. In *Proc. ICASSP'97*, pp. 719-722, Munich.
25. S. Wrigley, G. Brown, V. Wan, and S. Renals (2005). Speech and crosstalk detection in multichannel audio. IEEE Trans. Speech& Audio Proc., 13(1):84–91.