Transcription of Conference Room Meetings: an Investigation

Thomas Hain⋆ , John Dines† , Giulia Garau‡ , Martin Karafiat◁, Darren Moore†
Vincent Wan⋆, Roeland Ordelman ≀, Steve Renals‡

Department of Computer Science⋆
University of Sheffield
Sheffield S1 4DP, UK

IDIAP†
Martigny
Switzerland

Centre for Speech Technology Research‡
University of Edinburgh
Edinburgh EH8 9LW, UK

Faculty of Information Technology◁
Brno University of Technology
Božetěchova 2, Brno, 612 66, Czech Republic

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. In this paper we explore the use of various meeting corpora for the purpose of automatic speech recognition. In particular we investigate the similarity of these resources and how to efficiently use them in the construction of a meeting transcription system. The analysis shows distinctive features for each resource. However the benefit in pooling data and hence the similarity seems sufficient to speak of a generic "conference meeting domain". In this context this paper also presents work on development for the AMI meeting transcription system, a joint effort by seven sites working on the AMI (augmented multi-party interaction) project.

1. Introduction

The focus in large vocabulary automatic speech recognition research has been devoted to the transcription of speech found in natural environments for quite some time. The recorded speech is rarely planned but spontaneous or even conversational which contributes to relatively poor performance on these tasks. More recently more attention was devoted to the automatic transcription of conference room meetings. This interest is partly driven by the direct demand for transcripts of meetings. Moreover these transcripts can form the basis for higher level processing such as content analysis, summarisation, analysis of dialogue structure etc. This increased interest is manifest in yearly evaluations of speech recognition systems by NIST [7] or the existence of large scale projects such as AMI [11]. Initial work on meeting transcription was facilitated by the collection of the ICSI meeting corpus and the NIST meeting transcription evaluations in 2002. Further meeting resources were made available by NIST [10] and Interactive System Labs (ISL) [1] prior to the 2004 NIST RT04s Meeting evaluations [7].

As work in this domain is new many questions relating to fundamental properties of the data are yet unanswered. It is evident that the data varies greatly with the acoustic environment, the recording conditions and the content. A variety of recording configurations using either distant or speaker associated microphones poses additional challenges. Overlapped speech or reverberation in the meeting room are a further cause degradation in recognition performance. As the type of meeting can vary formal to informal, and from discussion to presentations it is not clear if we can speak of a general meeting domain. In this paper we investigate properties of data from several different sources with the aim to understand differences between them.

Our work is based on the AMI speech recognition system under development at many sites participating in the AMI project [11]. As the number of speech resources for meetings is still relatively small, similar to work presented in [9], a recognition system for conversational telephone speech (CTS) forms the starting point for our work on meetings. In the following we give a short 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.

1.1. Meeting resources

The ICSI Meeting corpus [8] 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 table-top microphones. A 3.5 hour subset of this corpus covering 7 meetings was set aside for testing (icsidev).Further meeting training data is available from NIST and ISL, with 13 and 10 hours respectively. Both NIST and ISL meetings are relatively free in their content (e.g. people playing games or discussing sales issues) and number of participants. In this work we make use of the RT04s NIST evaluation set (rt04seval) which contains data from the above as well as meetings recorded by the LDC.

A large collection and transcription effort of meetings is currently in progress as part of the AMI project [11]. Here meetings are collected at 3 different sites with a target corpus size of 100 hours of speech. Each meeting normally has four participants and a significant subset of meetings are task oriented. Each speaker wears headmounted and lapel microphones. As the recording and transcription of this corpus is still underway only 8 hours of limited quality transcriptions (amitrain05a) from task oriented meetings and one site are available for training. An additional development set (amidev) consisting of 8 meetings from 2 locations is used for testing.

2. The AMI CTS system

AMI recognition systems use standard technology such as HMM based acoustic models and N-gram based language models. For acoustic modelling HTK [12] or extensions thereof are used. Front-ends use 12 MF-PLP coefficients and c_0 for representing the speech signal. First and second order derivatives are added to form a 39 dimensional feature vector. Cepstral mean and variance normalisation is performed on complete channels.

[1] These corpora are available from the Linguistic Data Consortium (LDC).

Corpus	#words (MW)
Swbd/CHE	3.5
Fisher	10.5
Web (Swbd)	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 1: Size of various text corpora in million words (MW).

2.1. Dictionaries

The UNISYN pronunciation lexicon [1] 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 [2] with the CART system trained on the base dictionary. A left and right context of five letters as well as a left context of two phonemes was used. This gave 98% phone and 89% word accuracy on the base dictionary., for 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.

2.2. 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, HUB4-LM96 and Enron data (see below).

2.3. Language modelling

Language modelling data for conversational speech is sparse. Hence language models are constructed from other sources and interpolated (as in e.g. [3]). 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 the source would go beyond the scope of this paper. The most important non-standard data was found to be the the Web collected resources [14] and ICSI meetings [8]. 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 [13] 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.

Hub5e 1998/2001 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

Table 2: Perplexities on several NIST Hub5E evaluation test sets.

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

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

2.4. Acoustic modelling and adaptation

Acoustic models are phonetic decision tree state clustered triphone models with standard left-to-right 3-state topology were trained using standard HTK procedures [3]. Each state is represented as a mixture of 16 Gaussians. Smoothed heteroscedastic linear discriminant analysis (SHLDA) [5] is used to reduce a 52 dimensional (standard vector plus third derivatives) to 39 dimensions. Speaker adaptation is performed using vocal tract length normalisation (VTLN) both in training and test. Warp factors are estimated using a maximum likelihood criterion[3]. For further adaptation MLLR is used to transform means and variances[4].

2.5. Decoding and overall system performance

Decoding operates in multiple passes. The Cambridge University speech decoder HDecode is used for recognition with trigram language models. Table 3 shows results for systems trained on approximately 300 hours of Switchboard and Callhome data without or with SHLDA. The systems first generate output with non-VTLN models for use in VTLN warp factors estimation. The output of a second VTLN decoding stage is used for global MLLR (one transform for speech and one for silence) adaptation. 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.

3. Meeting resource analysis

Meeting data differs substantially from CTS. First 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.

3.1. Vocabulary

For the purpose of this paper we shall loosely define a domain as a set of sub-corpora that, when used in a combined

	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

Table 4: Statistics for meeting corpora.

	Vocabulary Source			
Corpus	ICSI	NIST	ISL	AMI
ICSI	0.00	4.95	7.11	6.83
NIST	4.50	0.00	6.50	6.88
ISL	5.12	5.92	0.00	6.68
AMI	4.47	4.39	5.41	0.00
ALL	1.60	4.35	6.15	5.98

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

	Vocabulary Source			
Domain	ICSI	NIST	ISL	AMI
ICSI	0.01	0.47	0.58	0.57
NIST	0.43	0.09	0.59	0.66
ISL	0.41	0.37	0.03	0.57
AMI	0.53	0.53	0.58	0.30
ALL	0.16	0.42	0.53	0.55

Table 6: %OOV rates of padded vocabularies. Columns denote the word list source, rows the test domain.

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 Out Of Vocabulary (OOV) rates using vocabulary derived from each meeting corpus. The OOV rates do not correlate perfectly with vocabulary sizes (Table 4). Overall the mismatch of ISL vocabulary to the other corpora is greatest. Table 6 shows the same analysis as before, however in this case the wordlists are padded as described in section 2.2. 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.

3.2. 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 were obtained by optimisation of interpolation weights for the components outlined in Table 1. Table 7 shows the weights. Note that both ICSI and AMI data show a strong bias towards their own source. Even though relevant for vocabulary selection, Broadcast News material appears to be of little importance. Table 8 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.

4. Meeting transcription

As meeting resources are still sparse it is necessary to find the appropriate background source material for acoustic model training. Systems at the NIST RT04s evaluations made use of either Broadcast News or CTS systems for bootstrapping. Practical evidence and the results in Table 7 suggest that CTS data is closer to this task. As CTS data is only available at a bandwidth of 4kHz this poses additional questions on the initialisation and training procedure .

Table 9 shows recognition performance on the icsidev test

	Optimisation target			
Corpus	ICSI	NIST	ISL	AMI
AMI	-	-	-	0.40
ISL	-	-	0 .18	-
NIST	-	0.15	-	-
ICSI	0.42	-	-	-
Web (meetings)	0.20	0.30	0.30	0.14
Switchboard	0.15	0.16	0.12	0.08
Fisher	0.12	0.18	0.22	0.16
Web (Swbd)	0.03	0.07	0.05	0.04
Web (fisher)			0.03	
Web (fisher topics)	0.06	0.14	0.11	0.15
HUB4-LM96	0.03			0.03

Table 7: Interpolation weights for trigram models and optimised perplexities on rt04seval and amidev (i.e. the corresponding subsets). ”-” denotes a-priori exclusion.

Corpus	ICSI	NIST	ISL	AMI	ALL
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
ALL	107.46	105.93	105.73	90.62	92.74

Table 8: Cross meeting room perplexities on the various meeting room specific eval sets from rt04seval and amidev. ALL denotes training or testing using all meeting data.

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 [6]. As CTS is NB only, adaptation to MN ICSI data was performed. An iterative application of MAP adaptation was found to give better performance. However the performance of the adapted system was still poorer than that of the system trained on WB data.

The above results show that MAP adaptation from CTS models while using wideband data is desirable. For MAP the adaptation model set is used for two purposes: for computation of state level posteriors and to serve as a prior. 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. After an initial step with MLLR-adapted CTS models iterative MAP adaptation is resumed as before. After 8 iterations a further 0.9% reduction in WER is obtained.

4.1. Meeting resource specific modelling

Similar to experiments on vocabulary and language models we are interested in the similarity of acoustic data for different corpora. Hence MRS acoustic models have been trained and tested in conjunction with the MRS language models and dictionaries. From the experiments presented above it is clear that adaptation of CTS models yields good performance even in the case of relatively large amounts of training data. We assume that is also true if the amount of meeting data is small. Hence all meeting room specific models were trained using MLLR-adapted CTS models in MAP adaptation to the specific meeting corpus. Table 10 shows WER results using MRS models as well as mod-

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

System	AMI	ISL	ICSI	NIST	TOT
MRS-AMI	53.8	63.7	52.3	59.0	56.7
MRS-ISL	54.9	57.4	48.0	53.3	55.4
MRS-ICSI	43.9	45.8	25.6	37.3	43.4
MRS-NIST	52.7	55.7	43.8	42.7	52.6
NOTAMI	40.9	45.7	25.1	34.3	40.9
ALL	40.0	45.2	26.0	33.5	40.2
ADAPTALL	39.1	44.5	25.6	34.4	39.4

Table 10: %WER on amidev(AMI) and the rt04eval sets. TOT gives WERS on both test sets. ALL denotes training on all meeting data, NOTAMI on ALL but AMI data. ADAPTALL stands for adaptation of CTS models to ALL data.

els trained on increasing amounts of meeting data. An initial observation makes clear that the construction of MRS systems does not guarantee individual best performance, however this appears to be largely due to the imbalance in training data size. The overall performance is clearly inferior to that of systems trained on the complete corpora. By training on all meeting data we can reduce the overall WER to 40.2%. This can be further reduced by 0.8% using MLLR-MAP adaptation from CTS models. In order to normalise for test-data specific WER variations Table 11 shows the relative increase in WER with meeting corpus specific models. In each row the smallest difference is obtained on data from the source the models were trained on. This suggests that a bias remains in acoustic modelling. We also conducted the same experiments with unbiased language models and word lists and obtained similar results.

5. Conclusions

In this paper we have presented an investigation into the transcription of various meeting resources. In particular we have addressed questions of linguistic and acoustic overlap between 4 major meeting resources: ICSI, NIST, ISL, and a first portion of the new AMI corpus. As meeting speech appears to have strong similarities to CTS the 2004 AMI transcription system for CTS which forms the basis of our meeting data work, was presented. An analysis of vocabularies, language and acoustic models in the meeting domain was presented. In all cases distinctive features for each domain were found. While vocabulary differences could be lessened by general padding differences in language modelling cannot considered to be minor. Acoustic models benefit from pooling data and this was found to outweigh meeting resource specific modelling approaches.

	AMI	ISL	ICSI	NIST
MRS-AMI	**37.6**	43.1	104.3	71.5
MRS-ISL	40.4	**29.0**	87.5	54.9
MRS-ICSI	12.3	2.9	**0.0**	8.4
MRS-NIST	34.8	25.2	71.1	**24.1**

Table 11: %Relative increase in WER of MRS systems compared to the ADAPTALL system (Table 10).

6. Acknowledgements

This work was largely supported by AMI[11]. The authors thank the rest of the AMI-ASR team for their valuable contributions: Lukas Burget, Iain McCowan. Jan Czernocky, Mike Lincoln, Jithendra Vepa and Chuck Wooters. We further 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.

7. References

[1] Susan Fitt (2000). Documentation and user guide to UNISYN lexicon and post-lexical rules, Tech. Rep., Centre for Speech Technology Research, Edinburgh.

[2] A.W. Black, P. Taylor and R. Caley (2004). The Festival Speech Synthesis System, Version 1.95beta. CSTR, University of Edinburgh, Edinburgh.

[3] T. Hain, P. Woodland, T. Niesler, and E. Whittaker (1999). The 1998 HTK system for transcription of conversational telephone speech. *Proc. IEEE ICASSP*, 1999.

[4] M.J.F. Gales & P.C. Woodland (1996). Mean and Variance Adaptation within the MLLR Framework. *Computer Speech & Language*, Vol. 10, pp. 249–264.

[5] L. Burget (2004), Combination of Speech Features Using Smoothed Heteroscedastic Linear Discriminant Analysis. in Proc. ICSLP'04, Jeju island, KR, 2004, p. 4.

[6] J.-L/ Gauvain, C. Lee (1994). MAP estimation for multivariate Gaussian mixture observation of Markov Chains, *IEEE Trans. Speech & Audio Processing*, 2, pp. 291-298.

[7] Spring 2004 (RT04S) Rich Transcription Meeting Recognition Evaluation Plan. NIST, US. Available at `http://www.nist.gov/speech`.

[8] 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.

[9] A. Stolcke, C. Wooters, N. Mirghafori, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin, & M. Ostendorf (2004). Progress in Meeting Recognition: The ICSI-SRI-UW Spring 2004 Evaluation System. NIST RT04 Meeting Recognition Workshop, Montreal.

[10] J.S. Garofolo, C.D. Laprun, M. Michel, V.M. Stanford, E. Tabassi (2004). The NIST Meeting Room Pilot Corpus, in Proc LREC'04.

[11] Augmented Multi-party Interaction. EC Project No. 506811, `http://www.amiproject.org`.

[12] The Hidden Markov Model Toolkit (HTK). `http://htk.eng.cam.ac.uk`, Cambridge University, UK.

[13] The SRI Language Modelling Toolkit (SRILM). `http://www.speech.sri.com/projects/srilm`, SRI international, California.

[14] 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.