DEVELOPMENT OF THE CU-HTK 2004 BROADCAST NEWS TRANSCRIPTION SYSTEMS

D.Y. Kim, H.Y. Chan, G. Evermann, M.J.F. Gales, D. Mrva, K.C. Sim, P.C. Woodland

Engineering Department, Cambridge University, Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {dyk21,hyc27,ge204,mjfg,dm312,kcs23,pcw}@eng.cam.ac.uk

ABSTRACT

This paper describes our recent work on improving broadcast news transcription and presents details of the CU-HTK Broadcast News English (BN-E) transcription system for the DARPA/NIST Rich Transcription 2004 Speech-to-Text (RT04) evaluation. A key focus has been building a system using an order of magnitude more acoustic training data than we have previously attempted. We have also investigated a range of techniques to improve both Minimum Phone Error (MPE) training and the efficient creation of MPE-based narrow-band models. The paper describes two alternative system structures that run in under 10×RT and a further system that runs in less than 1×RT. This final system gives lower word error rates than our 2003 system that ran in 10×RT.

1. INTRODUCTION

The accurate automatic transcription of broadcast material remains a challenging problem. One approach to improving accuracy is to greatly increase the amount of training data. The use of lightly supervised training [5, 1, 6] can yield a large increase in training data volume at low cost by using audio sources for which only closed captions exist. This has led to an order of magnitude increase in training data for the US English broadcast news (BN-E) task.

In this paper, we discuss how we have used up to 1350 hours of acoustic training data. The effect of increases in training data size is analysed in terms of word error rate (WER) reduction as more data is included in training. We have continued to use the discriminative lightly supervised training approach presented in [1]. We have also investigated a number of changes to the overall training procedure and evaluated these both individually and in the context of complete systems.

The paper is arranged as follows. First, an overview our 2003 BN-E system is given and then recent improvements in training acoustic models is presented. This includes the use of a dynamic maximum mutual information (MMI) prior in minimum phone error (MPE) training; an efficient method for building discriminative narrow-band models, and performance improvements using increased training data. We then give a description of two less than 10×RT systems and a 1×RT system which were developed for the RT04 evaluation.

2. CUED RT03 BN-E SYSTEM

The system, developed for the March 2003 Rich Transcription (RT03) evaluation, runs in a little under 10×RT. It operates in multiple passes and includes multiple branches that use different acoustic models. In the final stage, system combination is used to combine the outputs from the separate branches. Full details of the system structure and the models involved are given in [4].

Each frame of speech is represented by 13 PLP coefficients with first, second and third derivatives appended and then projected down to 39 dimensions using HLDA. The cross-word triphone HMMs, which use 7000 clustered states each with a 16 component Gaussian mixture distribution, were estimated using the BN-E data released by the LDC in 1997 and 1998 (`bnac`). Since some BN data, for example telephone interviews, are transmitted over bandwidth-limited channels, both wide-band and narrow-band spectral analysis variants of each model set were trained. All model sets were trained using MPE [7] and gender-dependent versions were derived using MPE-MAP [8]. A number of broadcast and newswire text corpora were used to train a word 4-gram language model (LM).The overall system decoding structure is as follows:

P1 initial transcription: The P1 pass (gender independent models) provides an initial word-level transcription which is used for both gender determination and as the adaptation supervision for the P2 models.

P2 lattice generation: The adaptation uses global least squares regression mean transforms and MLLR variance transforms. Word lattices are generated using the adapted acoustic models and a 4-gram word LM.

P3 lattice rescoring: Two separate model sets are used to rescore the P2 lattices. The P3.1 system was built using speaker adaptive training (SAT) employing global constrained MLLR transforms. The P3.2 system was trained in the normal speaker-independent fashion but employed a special single pronunciation (SPRON) dictionary. Both P3 model sets were adapted using lattice MLLR adaptation and a global full-variance transform.

Each of the stages P2, P3.1 and P3.2 produce word lattices and these are converted to confusion networks and then combined with confusion network combination (CNC) [2]. Finally, a forced alignment of the final word-level output was used to obtain accurate word times before scoring. The full system ran in 9.1×RT[1] on the 2003 evaluation set (`eval03`).

3. TRAINING AND TEST DATA SETS

In addition to the `bnac` set used in the RT03 system, three more training data sets were used for acoustic model training. The `tdt4` corpus was originally developed for the topic detection and tracking task and closed caption text is available. Recently, the LDC provided more broadcast news data, which we denote `tdt4a` and `bn03`, so that the total amount of training data has been greatly increased. A brief summary of these data sets is given in Table 1.

This work was supported by DARPA grant MDA972-02-1-0013. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

[1] On a single Intel Xeon 2.8GHz/512kB L2 cache processor.

data	description	size (hours)	
		original	usable
bnac	RT03 training data	–	143
tdt4	TV+radio / 6 src / Oct00-Jan01	300	231
tdt4a	TV / 4 src / Mar01-Jul01	530	377
bn03	TV / 19 src / Mar03-Nov03	6375	–

Table 1. Available BN-E training data and size.

As no detailed transcriptions (only closed captions) are available for tdt4 and tdt4a data, we used lightly supervised training [1]. This technique performs a recognition pass on the training data with a data set specific biased LM which is trained in part on the closed caption data. The recognised output is then used as the transcription for training. The biased LM is constructed by building an LM on just the closed caption text and then interpolating this with a more general LM (RT03 LM) which includes detailed broadcast news transcriptions. The recognition system is a simplified version of the RT03 BN-E system which includes only the P1 and P2 stages and the final output is formed by confusion network decoding of the 4-gram lattice output from the P2 stage. This system normally runs within 5×RT and we call it the P1-P2 system. As we normally automatically remove advertisements and then remove additional audio during the segmentation process, the amount of usable data is reduced from the length of the original audio files. Furthermore, we excluded tdt4 data after 15 January 2001 so not to have temporal overlap of acoustic or language model data with the development test sets.

Transcriptions and segmentations of the bn03 data was provided by BBN using a recognition/filtering approach and the detailed method was presented in [6]. We selected two subsets from BBN's transcription for bn03, each of about 300 hours. The first set was sampled from six major sources included in the data from ABC, CNBC, CNN, CNNHL, CSPAN and PBS. The second cset omes from CNN and the other six sources (CBS, FOX, MSN, MSNBC, NBC, NWI) which were not included in the first set. We included no data in these sets that was later than 14th November 2003, so that temporal overlap with development data was avoided.

training set	description	size(hours)
bntr04-base	bnac+tdt4	375
bntr04-750h	+tdt4a	752
bntr04-1050h	+1st selection of bn03	1050
bntr04-1350h	+2nd selection of bn03	1350

Table 2. Selected BN-E training data sets and sizes.

By adding the additional training data incrementally, we have 4 different training sets as given in Table 2. The acoustic models were initially built using either the bntr04-base or the bntr04-750h data. Later more parameters were added to the models and trained with the bntr04-1050h and bntr04-1350h data sets. These final model sets were actually used in our RT04 evaluation systems.

3.1. Test Data

For development we used two data sets from the TDT4 sources broadcast in late January 2001 (dev03 and dev04) as well as the RT03 evaluation set from February 2001 eval03. We also used the LDC-released set of development data set, denoted here (dev04f) which was broadcast in late November 2003[2]. The dev04f data is from different sources to the earlier data and contains a larger amount of more difficult data (such as high levels of background noise/music and non-native speakers) than other test sets.

4. REVISED ACOUSTIC MODELLING

4.1. MPE Training with Dynamic MMI prior

Use of the I-smoothing technique is necessary to allow good generalisation performance in MPE training [7]. I-smoothing uses a more robust estimate of model parameters as a prior distribution in MPE training. Hence the prior parameters of the I-smoothing distribution act as back-off value to the MPE estimate. In standard MPE training, maximum likelihood (ML) estimates of mean and variance are used as the priors. As the priors are generated for each iteration based on the current model parameters, they are referred to as a *dynamic* prior.

As much more data per parameter becomes available for MPE training, it is possible to robustly estimate a more appropriate prior distribution such as one based on an MMI estimate [9] and this can be used in place of (or in addition to) the normal ML prior estimate. To implement this change, since ML statistics (equivalent to MMI numerator statistics) are already being used in MPE training, only MMI denominator statistics are additionally required to use a dynamic MMI prior. Also there is no extra computation in accumulating statistics compared to MPE training using ML prior, since the MMI denominator uses the same component posterior occupancy generated by lattice forward-backward algorithm in standard MPE training.

A comparison between a dynamic MMI prior and a dynamic ML prior is given in Table 3. It can be seen that a reduction in WER of 0.3% and 0.1% abs for dev03 and eval03 respectively is obtained by using the dynamic MMI prior. For this size of training set we didn't find it to be necessary to include an ML I-smoothing term in the MMI prior estimate.

Training	%WER	
	dev03	eval03
MPE (ML prior)	13.9	12.6
+GD MPE-MAP	13.7	12.4
MPE (MMI prior)	13.6	12.5
+GD GI-MPE prior	13.5	12.3

Table 3. %WER for (GI) MPE model with dynamic MMI prior and GD MPE models using bntr04-base. 16 comp/state. Single pass decoding with the RT03 trigram LM. NB segments decoded using the RT03 MPE NB models.

4.2. Gender Dependent (GD) MPE training

Normally we build GD models using a few additional iterations starting from a GI model set. Of course care must be taken to avoid over-training and typically we only update the Gaussian means and mixture weight parameters. Furthermore it is useful to use more conservative prior parameter settings. Therefore for GD MPE training we have used the parameters of the GI MPE model as the I-smoothing priors throughout the training of the GD model set.

[2]One show was broadcast on 1st December 2003.

This is referred to as a *static* prior in contrast to the *dynamic* prior in GI MPE training. As the static prior is not updated during GD MPE training, it is more robust than the dynamic prior. The GD MPE models showed consistent gain over GI models as shown in Table 3.

4.3. Narrow-band Models

In the CUED RT03 BN-E system, HMM sets for narrow-band (NB) data were built during the ML training stage then the MPE training was performed independently on NB filtered data. However as the amount of training data increases, the cost of another full set of MPE training iterations to build NB models is too high. To alleviate this problem, the single pass retraining method was extended to work with the MPE criteria and then directly applied to build NB model using the WB MPE model.

Training Method	Iter	%WER		
		dev03	eval03	dev04
NB MPE	8	14.9	13.6	16.5
MPE-SPR (ML prior)	—	15.0	13.8	16.6
+MPE	1	14.7	13.7	16.4

Table 4. % WER of various bnac NB acoustic models. WB segments hypothesis using the RT03 WB MPE model.

Table 4 presents the performance of NB models trained using various methods. The MPE-SPR with ML priors showed slightly poorer performance than the full MPE trained NB models, but another iteration of MPE training (with ML prior) gave 0.2% and 0.1% absolute gain on dev03 and dev04 respectively, and was only 0.1% worse on eval03 than the RT03 NB model.[3]

4.4. Increased Training Data/Model Complexity

As the quantity of available training data has increased by a large amount, we investigated also increasing the number of parameters in the HMMs by both increasing the number of Gaussians per state and the number of clustered states. The model structure for the basic HMMs used in our RT03 BN-E system included an average of 16 Gaussian mixture components per clustered state and 112k Gaussians in total (7k states).

We first doubled the average number of Gaussians per state to 32. With the bntr04-750h training set, there were reductions in WER of 0.6% and 0.3% for dev03 and eval03 as shown in Table 5. This is in contrast to our previous observations on increasing model complexity, for MPE models with bnac 143 hours training set, where we found little advantage to doing so.

We then increased the number of clustered states from 7k to 9k and also switched training set to bntr04-1050h. The results still showed consistent gains over various test sets. Finally we used the bntr04-1350h set for MPE training. There were further 0.1-0.2% gains for dev03 and eval03 and another larger change of 0.7% abs for the dev04f set. Note that this relatively big gain in dev04f is partly due to similarity between dev04f and bn03 training set.

[3]This NB model building method was evaluated later using P1-P2 system with the RT04 LM, and found that the new NB model with additional training data consistently outperformed the RT03 NB models.

Training Data		%WER		
		dev03	eval03	dev04f
bntr04-750h	16/7k	13.4	12.1	–
bntr04-750h	32/7k	12.8	11.8	21.6
bntr04-1050h	32/9k	12.2	11.4	20.3
bntr04-1350h	32/9k	12.1	11.2	19.6

Table 5. %WER with GI MPE models with different training sets. Single pass decoding of WB segments with the RT03 trigram LM. NB segments decoded using the RT03 NB MPE model.

4.5. SPRON

All the acoustic models discussed so far have been based on a multiple pronunciation per word (MPRON) dictionary. As in our RT03 system, we also built model sets based on a single pronunciation (SPRON) dictionary. The SPRON dictionary was generated based on pronunciation statistics from bntr04-1050h training set.The same training procedure including the dynamic MMI prior and GD MPE training was performed to build SPRON acoustic models. The performance of bntr04-1350h SPRON GD MPE models was 10.8% and 12.6% WER for eval03 and dev04, which is 0.2-0.3% abs better than the MPRON counterparts with the same experimental setup given in Table 5.

5. RT04 EVALUATION SYSTEM

5.1. Language Model

The LM for RT04 was built in a similar way as RT03 BN-E LM in which five 4-gram models were linearly interpolated together. Small to mid-size models were smoothed using modified Kneser-Kney discounting and components trained on large data sets used Good-Turing discounting. After the interpolation, the component models were merged and the final model was pruned with entropy-based pruning.

The main difference in the R04 LM was the addition of the new training texts. The closed caption text from tdt4a and bn03 were added to the text corpus. Also more transcriptions from CNN's website and various newswire texts were newly added. As a result, the number of words for LM building was increased by 40% compared to RT03 LM, i.e. 1.4 billion word tokens were used in total. The interpolation weights were optimised on a text comprised of eval03, dev04, and dev04f. The perplexity values with the RT04 4-gram LM were 120, 118, and 132 for eval03, dev04 and dev04f.

5.2. Rover Using Dual Segmentations

Although the performance of automatic audio segmentation systems is generally good, there are still places where a particular segmenter makes errors. Therefore a strategy was investigated in which two independent segmentations were used and separate recognition systems run on each and the final word-level outputs combined. This method is more robust in the sense that when one segmenter fails in a particular acoustic environment, the other may work well or vice versa.

To implement this approach we ran the P1-P2 system with both the CUED RT03 and LIMSI's segmentation [3] independently, then performed ROVER using the two decoding results. As shown in Table 6, the performance after ROVER is consistently better than best single system on three different test sets, and the gain was 0.3%-0.4% abs.

Segment	%WER		
	eval03	dev04	dev04f
CUED	9.2	11.9	16.6
LIMSI	8.8	11.4	16.2
ROVER	8.5	11.0	15.8

Table 6. %WER of P1-P2 system and ROVER using CUED and LIMSI segmentations. bntr04-1050h WB models, the RT03 NB models. RT04 LM.

5.3. 10×RT System

First an RT03 style system was built with the new acoustic/language models and using the LIMSI segmenter. The SAT model was trained on bntr04-1050h in the same way as RT03 system, apart from using automatically generated speaker clustering for the new training data (tdt4, tdt4a and bn03). The GD SPRON models were trained using bntr04-1350h. The system ran in 8.4×RT[4] on the RT04 evaluation data set and on average resulted in a 22% relative reduction in WER.

We built an alternative 10×RT system (RT04 10×RT(2)) using the dual segmentation approach with a simplified system to reduce the run-time of each system to under 5×RT. These changes included (1) using a faster decoding configuration in P1 with no 4-gram expansion (0.3-0.4×RT); (2) slightly narrower beamwidth in P2 ($<$ 3×RT); (3) single P3 branch using SPRON model (1.5×RT); (4) CNC using P2 and P3. The SPRON model was chosen as a single P3 branch since it showed better performance than the SAT in the RT03 style system. The performance of this system is given in Table 7. The final ROVER combined numbers show 0.2-0.4% abs lower WER than the RT03 structure.

System		%WER		
		eval03	dev04	dev04f
RT03 10×		10.6	13.2	18.6
RT04 10×(1)	RT03 style	8.0	10.4	14.9
RT04 10×(2)	CUED-seg	8.4	10.8	15.5
	LIMSI-seg	8.1	10.3	14.9
	ROVER	7.8	10.0	14.7

Table 7. %WER of the RT04 10×RT systems

5.4. 1×RT System

A 1×RT system was built that has a very similar structure to the P1-P2 system described above. All the decoding configurations were carefully tuned to make the system faster while minimising the performance degration relative to the 10×RT system. In particular the performance of the the P2 stage was relatively insensitive to the WER of the first stage.

The 1×RT system consists of (1) LIMSI segmentation (0.15×RT) (2) a very fast P1 using the MPE WB model with 16 Gaussian mixture components and a small sized trigram LM (0.15×RT), and no 4-gram expansion; (3) unsupervised adaptation and a fast P2 decoding with trigram (0.6×RT) with 4-gram expansion; (4) confusion network decoding and forced alignment.

In the final stage, word tokens with low confidence scores were removed from the recognised results. As shown in Table 8, the very fast P1 stage still produces reasonable error rates and the final output gives lower WERs than the RT03 10×RT system for all data sets.

[4]All times for RT04 systems were run on an Intel Xeon 3.2GHz/2MB L3 cache processor with hyperthreading enabled.

Pass	%WER		
	eval03	dev04	dev04f
P1	17.2	21.7	27.8
P2	9.9	12.7	17.4
final	9.8	12.5	17.3

Table 8. %WER of the RT04 1×RT system for development sets. final is after removing low confidence words.

6. CONCLUSIONS

We presented the CU-HTK 2004 BN-E systems for the DARPA/NIST RT04 evaluation. Using improved acoustic/language models and by combining systems with different segmentations, the 10×RT system gave on average a 24% relative reduction in WER over the RT03 system. We also presented a 1×RT system which outperforms the RT03 10×RT system.

7. ACKNOWLEDGEMENT

BBN provided the bn03 transcriptions and LIMSI segmentations that were used for this work. We would like to thank all of the HTK STT team, in particular K. Yu, L. Wang and X. Liu.

8. REFERENCES

[1] H.Y. Chan & P.C. Woodland, "Improving Broadcast News Transcription by Lightly Supervised Discriminative Training," *Proc. ICASSP*, 2004.

[2] G. Evermann & P.C. Woodland, "Posterior Probability Decoding, Confidence Estimation and System Combination." *Proc. Speech Transcription Workshop*, 2000.

[3] J. L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast Transcription Systems," *Speech Communication*, vol. 37, pp. 89–108, 2002.

[4] D.Y. Kim, G. Evermann, T. Hain, D. Mrva, S.E. Tranter, L. Wang, & P.C.Woodland, "Recent Advances in Broadcast News Transcription," *Proc. ASRU*, 2003.

[5] L. Lamel, J. L. Gauvain & G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

[6] L. Nguyen & B. Xiang, "Light Supervision in Acoustic Model Training," *Proc. ICASSP*, 2004.

[7] D. Povey & P.C. Woodland, "Minimum Phone Error and I-Smoothing for Improved Discriminative Training," *Proc.ICASSP*, 2002.

[8] D. Povey, M.J.F. Gales, D.Y. Kim & P.C, Woodland "MMI-MAP and MPE-MAP for Acoustic Model Adaptation," *Proc. Eurospeech*, 2003.

[9] G. Saon, D. Povey & G. Zweig, "CTS Decoding Improvements at IBM," *EARS STT Workshop*, St. Thomas 2003.