TRAINING LVCSR SYSTEMS ON THOUSANDS OF HOURS OF DATA

G. Evermann, H.Y. Chan, M.J.F. Gales, B. Jia, D. Mrva, P.C. Woodland, K. Yu*

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, UK
Email: ge204@eng.cam.ac.uk

ABSTRACT

Typical systems for large vocabulary conversational speech recognition (LVCSR) have been trained on a few hundred hours of carefully transcribed acoustic training data. This paper describes an LVCSR system for the conversational telephone speech (CTS) task trained on more than 2000 hours of data for which only approximate transcriptions were available. The challenges of dealing which such a large data set and the accuracy improvements over the small baseline system are discussed. The effect on both acoustic and language modelling performance is studied. Overall increasing the training data size from 360h to 2200h and optimising the training procedure reduced the word error rate on the DARPA/NIST 2003 eval set by about 20% relative.

1. INTRODUCTION

Conversational Telephone Speech (CTS, formerly known as Hub5 or Switchboard) is still a hard task for automatic speech recognition. Over the last decade there has been a fairly consistent research effort by a small number of groups funded by the U.S. government. The progress in system performance was driven and monitored by yearly evaluations conducted by DARPA/NIST. The typical systems entered in these evaluations increased in complexity significantly and steady progress was made (up to about 10% relative word error rate reduction per year). However, the problem is far from solved and many researchers felt a more aggressive approach was needed to accelerate progress. The DARPA EARS program that began in 2002 offered the opportunity to investigate more aggressive approaches.

In EARS the main Speech-to-Text contractors (BBN, CUED, LIMSI, SRI, IBM) decided that a massive increase of the available acoustic training data would offer the best chance for achieving significant performance improvements in a short amount of time.

For the last decade the CTS training set consisted of 200-300 hours of carefully transcribed data. It was decided to collect thousands of hours of new data and to experiment with new strategies to reduce the cost of the manual transcription of the training data.

This paper describes experiments with the first batch of data that became available in the middle of 2004. This set contained about 2000 hours of telephone conversation data. The starting point of the experiments was the Cambridge University CTS system described in [1] which was trained on 360h of acoustic data.

*Gunnar Evermann is now with VoiceSignal Technologies http://www.voicesignal.com

2. CTS TASK & DATA RESOURCES

The CTS data consists of phone conversations between volunteers on an assigned topic in (American) English. To rapidly increase the amount of available training data the LDC implemented a new collection protocol ("Fisher" collection). A detailed discussion of this new protocol can be found in [2]. About 2000h of new Fisher data was collected over the last year.

Since the cost of creating careful manual transcriptions would have been prohibitive it was decided to rely on "quick transcriptions", which were generated by a commercial transcription service and post-processed automatically. BBN employed the transcription service WordWave to produce transcriptions for about 1800h while the LDC produced quick transcriptions for the remaining 200h. An overview of the respective procedures can be found in [3] and [4], respectively.

The following data sets were available for training and testing the CUHTK CTS system:

h5train03b 360h data set used in 2003 evaluation

fisher3896 520h Fisher data set, 3896 conversations with WordWave quick transcriptions (early version)

fsh2004 1820h, BBN/WordWave + LDC quick transcriptions

fsh2004h5train03b 2180h all available CTS data.

eval03 test 6h set. Fisher and Swb2-5 data, 72 conversations

dev04 test 3h set Fisher, 36 conversations

3. DATA PREPARATION

As usual a certain amount of data cleanup had to be performed on the audio data and associated transcriptions that were provided for the Fisher 2004 data. The original transcriptions consisted of 1940h data (1758h BBN data, 182h LDC data). The text was processed to normalise spelling and transcription conventions. About 11.000 replacement rules were generated for this purpose (the majority of these were related to word fragments). Pronunciations for about 6800 words that occurred at least twice were added to the CUHTK training dictionaries. About 18h worth of segments containing new words that only occurred once were discarded.

Forced alignment was performed on all of the new data and based on this errorful segments were discarded (<30h) and the length of silence portions at segment boundaries was normalised. After this processing 1819h of data remained, 1042h from female speakers and 777h from male speakers.

4. TRAINING ON QUICK TRANSCRIPTIONS

In order to investigate the impact of using quick transcriptions for acoustic modelling, a set of experiments was performed based on 20 hours of Switchboard 1 data for which four different sets of transcriptions were available. The first set were the Mississippi State University careful transcriptions (MSU). The second set were LDC quick transcriptions. The third set and the final set were provided by BBN and used different post-processing strategies of the WordWave quick transcriptions ("Algorithm I" and "Algorithm III", respectively) [3]. These two BBN generated transcriptions differed in the way that words were assigned to segments and in the quality checking procedure. Acoustic models were trained using Maximum Likelihood (ML) and Minimum Phone Error (MPE) estimation for each of these transcriptions.

	dev01		eval03	
	ML	MPE	ML	MPE
MSU	43.4	40.5	43.5	40.5
LDC QT	43.6	41.2	43.8	41.2
BBN WWave1	43.6	41.2	44.0	41.4
BBN WWave3	43.4	40.8	43.6	40.8

Table 1. %WER, unadapted, trigram, ML and MPE models

From table 1, it can be seen that using quick transcriptions causes an increase in word error rate. This can be explained by the larger number of transcription errors that occur in the quick transcriptions. Another observation is that MPE discriminative training is more sensitive to transcription quality than ML training. An encouraging result is that there is only a small performance gap between the models trained using the BBN "Algorithm III" Word Wave quick transcriptions and the MSU careful transcriptions. Based on these results and similar findings using BBN's ML training procedure it was decided to transcribe the whole new Fisher collection data using the quick transcription methodology.

5. LANGUAGE MODELLING

The Language Models used in the CUHTK systems are simple word-based n-gram models. Typically separate n-grams are trained on different text corpora and then interpolated together with the interpolation weights optimised on a development test set. The resulting LM is pruned using entropy-based pruning [5]. The components trained on smaller corpora (fewer than 20M words) are trained using Kneser-Ney discounting while Good-Turing discounting was used for the larger LMs.. The baseline language model in this paper is the interpolated fourgram model used in the 2003 CU-HTK system [1].

An additional component n-gram was built on the quick transcripts for the Fisher data. An additional data corpus was collected from the web by Bulyko & Ostendorf [6] by submitting frequent Fisher n-grams to Google as queries and normalising the returned pages. As out-of-domain data, transcripts and closed captions of broadcast news shows were used. Table 2 summarises the LM training set before and after updating the training texts.

As Table 3 shows, adding the highly relevant in-domain Fisher 2004 training data has a big impact on PP. Adding the new 529MW Web data decreased PP by another point. While experimenting with the new texts, it was found that the old (not Fisher-specific) 62MW Web data was redundant. Adding the additional out-of-domain data did not yield any perplexity improvement.

Training Text	Size (MW)	Weight
BN texts	427 → 488	0.23 → 0.05
cell1	0.2	0.11 → 0.02
che+swbdI	3.2	0.29 → 0.04
swbdII	0.9	0.23 → 0.05
Fisher 2004	0 → 21	0 → 0.67
Web	62 → 529	0.14 → 0.17

Table 2. Interpolation weights optimised on **dev04**. Training text size and weights before → after updating the training set.

Language Model	Weights optimised on	Perplexity
fgint03	dev01+eval00,01,02	62.0
fgint03	**dev04**	61.7
Fsh only	–	55.7
fgint03+Fsh	**dev04**	52.8
fgint03+Fsh+web	**dev04**	51.7
fgint03+Fsh+web+BN	**dev04**	51.7

Table 3. Perplexities of word 4-grams on **dev04**.

6. ACOUSTIC MODEL TRAINING STRATEGY

Due to the large amount of training it wasn't feasible to perform all experiments on the full data set. Instead a training strategy was adopted where initial experiments were conducted on subsets of the data.First of all the entire data set was pre-processed and prepared for use in experiments. This included cleaning up the transcriptions, aligning the data and determining VTLN warp factors. During this process a number of issues with the software and the general infrastructure were discovered and fixed (some of these issues are discussed in the following section).

A manageable subset of the data was selected for fast-turnaround experiments. Conversations were selected to yield a 400h subset (`fsh2004sub`) that was balanced for speakers' gender, conversation topics and had the same distribution of phone line conditions as the test data (25% cellular). Baseline ML and MPE models were built on this data subset to allow the investigation of advanced acoustic modelling techniques (see the companion paper [7] for details). Concurrently further models were built on larger data subsets (`fsh2004sub2`: 800h) and finally the whole data set (`fsh2004h5train03b`).

7. COMPUTATIONAL ISSUES

During the training setup a number of computational issues were encountered. The complete training set consists of:

- 2,180 hours (785 million frames)
- 30,660 conversation sides
- 1,803,682 segments
- 24.6 million words

Performing the acoustic training on this data set takes a large amount of compute time. For a standard triphone model set (9k state, 36mix) one ML training iteration takes 216 CPU hours (0.1xRT). The lattice generation for discriminative training takes more than 1 CPU year (4xRT) and one MPE iteration takes 880 CPU hours (0.4xRT). The training was performed on a compute cluster consisting of about 100 CPUs (2.4-3.2 GHz Pentium 4) running Linux.

Apart from the run time another challenge is the size of the data files involved in the training. To cope with the file server/network load a high-performance file server with Gigabit Ethernet (3.6 TB RAID5 storage) was used. The PLP feature files for full data set (2000+h stereo data) take 48GB. One set of lattices for discriminative training takes 89 GB. Each model set takes 100 MB.

8. ACOUSTIC MODELS

8.1. Baseline models

As a starting point and to verify the general setup baseline acoustic models were built on the 400h fsh2004sub data set. These triphone models use the same number of parameters as previous h5train03b models (6k tied states, 28 mixture components per state). Both ML and MPE models were trained.

From the `fsh04sub`results in table 4 it can be seen that new Fisher 400h set gives very similar performance to the old 520h set, which had lower transcription quality. Overall there was a WER reduction of 1% abs. over the 2003 training set.

To study the effect of using more data the training data size was increased incrementally. The number of parameters was not changed relative to the previous experiments. The results in table 4 show that while the ML models' performance only improves slightly when going beyond 400h of training data, the performance of the discriminatively trained models continues to improve substantially. Overall adding 1800h of Fisher to the acoustic training improves the eval03 WER for ML models by 1.5% abs. and for MPE models by 3.1% abs.

			eval03	03Sw	03Fi	dev04
ML	h5train03b	360h	31.7	36.1	27.1	28.1
ML	fisher3896	520h	30.8	34.7	26.6	26.9
ML	fsh04sub	400h	30.8	34.6	26.7	26.8
ML	fsh04sub2	800h	30.5	34.4	26.4	26.5
ML	fsh04h5t03b	2200h	30.2	34.1	26.0	26.4
MPE	h5train03b	360h	27.3	31.6	22.7	23.7
MPE	fisher3896	520h	26.2	30.0	22.2	22.3
MPE	fsh04sub	400h	25.9	29.6	21.9	21.9
MPE	fsh04sub2	800h	25.1	28.9	21.1	21.3
MPE	fsh04h5t03b	2200h	24.2	27.9	20.2	20.5

Table 4. %WER on eval03 and dev04, unadapted, 2003 trigram

8.2. Adapted models

To test these models with adaptation and the new LM the CUHTK 5xRT system ("P1-P2" system, see [8]) was rebuilt with the new models. In this system a very fast first pass (P1) produces an initial transcription that is used to perform VTLN and unsupervised speaker adaptation. In the second pass (P2) the adapted models are used to produce lattices on which confusion network decoding is performed.

From table 5 it can be seen that adding the Fisher data to the LM reduces the word error rate with last year's acoustic models by 1.3% abs. (1.6% on Fisher). Using 400h of Fisher data yield 0.6% lower WER than using last year's training data. Doubling the amount of Fisher data gives an additional 0.7%. The total WER reduction from adding Fisher data to the old 2003 data for acoustic and LM training is 3.3% abs. (2.9% on Fisher)

model		LM	eval03	03Sw	03Fi
h5train03b	360h	LM03	24.6	28.7	20.2
h5train03b	360h	LM03+fsh	23.3	27.6	18.6
fsh04sub	400h	LM03+fsh	22.7	26.7	18.4
fsh04sub2	800h	LM03+fsh	22.0	25.9	17.8
fsh04h5t03b	2200h	LM03 + fsh	21.3	25.1	17.3

Table 5. eval03 %WER, 5xRT P1-P2 system,MPE, word 4-gram

8.3. More Parameters

An obvious shortcoming in the above experiments is that the number of Gaussians in the acoustic model was kept fixed while the training data set was increased by a factor of six. Ways of increasing the number of parameters were investigated. The standard HTK model training procedure relies on iterative mixture splitting where the Gaussian with the highest mixture weight in each state is split. An alternative criterion was sought that would also take the variances of the Gaussians into account. The mixtures in each state were ranked separately based on the weights and the covariance determinant values. The Gaussian with the highest average rank was picked as a candidate for splitting.

System	Comp	eval03	eval03Sw	eval03Fi
Baseline	32	30.9	34.8	26.7
rank-based		30.7	34.6	26.5
Baseline	36	30.7	34.5	26.6
rank-based		30.4	34.4	26.1

Table 6. %WER on `eval03` for MLE fsh2004sub models with different mixup criteria

Table 6 compares the performance of the two mixing-up criteria. The training was started from a 16 component fsh2004sub system. The rank-based strategy outperforms the conventional approach and it was decided to build models on the full data set with 9k states and 36 mixture components using the rank-based criterion.

9. PRIORS IN MPE TRAINING

9.1. MMI priors for robust training

MPE training is susceptible to over-training. I-smoothing distribution, with a form of normal-wishart distribution as MAP, are then introduced to obtain robust MPE training and improve generalisation ability. The prior parameters of the I-smoothing distribution act as back-off values of the MPE estimate, which may significantly affect the performance of MPE training. The selection of priors is a key issue, in particular mean and variance priors. In standard MPE training, ML estimates of mean and variance are used as the priors. As the priors are generated on-the-fly in terms of sufficient statistics, they are referred to as *dynamic* priors. Considering that better back-off values may yield better performance, dynamic MMI priors may be used instead of the dynamic ML prior. For large amount of training data, the dynamic MMI prior is likely to be robust. To obtain dynamic MMI priors, ML statistics (equivalent to MMI numerator statistics) and

MMI denominator statistics are needed. A comparison between dynamic MMI prior to standard dynamic ML prior is shown in table 7 where MPE-τ^I is the manually selected value representing the equivalent"occupancy" of the MPE I-smoothing prior, while MMI-τ^I is for MMI I-smoothing prior. The dynamic MMI prior outperformed dynamic ML prior, especially on the fisher part of the dataset. Therefore, dynamic MMI prior were employed in all following experiments.

MPE Prior	MPE-τ^I	MMI-τ^I	eval03	03Sw	03Fi
Dyn. ML	50	—	26.3	29.9	22.5
Dyn. MMI	75	0	25.9	29.6	21.9

Table 7. %WER on eval03 for MPE models trained on fsh2004sub, unadapted, 2003 trigram

9.2. Gender-dependent MPE training

Gender-dependent MPE training is done on top of the well trained MPE-GI model. Since GD MPE training is even more susceptible over-training only mean and Gaussian mixture weights were updated and more conservative priors were used. Parameters of MPE-GI model were used as the I-smoothing priors, which can be referred to as the `static` prior. As the static prior is not updated during MPE training, it should be more robust than a dynamic prior.

Table 8 shows the performance of the GD MPE model compared with the GI MPE in the 5xRT system which includes adaptation.

System	MPE Prior	eval03	Male	Female
MPE-GI	Dynamic MMI	22.7	24.0	21.4
MPE-GD	MPE-GI model	22.4	23.8	21.0

Table 8. %WER on eval03, fsh2004sub models, adapted, LM03+Fsh 4-gram, P1-P2 system

10. OVERALL PERFORMANCE

All the changes discussed above were put together and the resulting models were tested in the framework of the CUHTK 5xRT system. Table 9 shows the performance on eval03 with adapted models, a fourgram LM and confusion network decoding.

		eval03	03Sw	03Fi
h5train03b(360h)	LM03	23.8	27.8	19.5
fsh2004h5t03b 6k	+fsh	20.7	24.5	16.7
fsh2004h5t03b.9k	+fsh	19.9	23.4	16.2
fsh2004h5t03b.9k.GD	+fsh	19.6	23.1	15.8
fsh2004h5t03b.9k.GD	+fsh+web	19.4	22.9	15.6

Table 9. %WER on eval03, adapted MPE models, word 4-gram+CN, 5xRT P1-P2 system

The results indicate that the use of more parameters in the acoustic model, gender dependent modelling and the inclusion of web data in the LM yields a further 1.3% abs. decrease in WER. Overall performance was improved by 4.4% abs. over last year's models. One of the concerns with training on large amounts of Fisher data which have a fairly limited set of topics is that the resulting models are tuned specifically for Fisher data. As a test the new Fisher-trained models were tested on the 2000 eval data set which contained both very hard Callhome and relatively easy Switchboard 1 data. The results are shown in table 10. The performance improvement is consistent across the subsets.

system	run time	WER	CHE	Swb1
2000 cuhtk1 eval	255xRT	25.4	31.4	19.3
2004 P1-P2 fsh04h5t03b	5xRT	19.6	24.5	14.8

Table 10. %WER on eval00 set (manual segmentation) with 2000 system and 2004 P1-P2 system

11. CONCLUSIONS

This paper discussed the issues in training LVCSR systems on large amounts of data. The creation of infrastructure for large scale experiments on the 2200 hour data set was described. Overall WER on eval03 was reduced by 4.4% abs. in a 5xRT system by adding 1800h Fisher data for acoustic and LM training and optimising the training procedures. From these performance numbers it is clear the the aggressive Fisher data collection combined with the quick transcription methodology was a very good investment

ACKNOWLEDGMENTS

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. The authors would like to thank all members of the CUED HTK STT team.

12. REFERENCES

[1] G. Evermann, H.Y. Chan, M.J.F. Gales, T. Hain, X. Liu, L. Wang D. Mrva, and P.C. Woodland, "Development of the 2003 CU-HTK conversational telephone speech transcription system," in *Proc. ICASSP*, 2004.

[2] C. Cieri, D. Miller, and K. Walker, "From switchboard to fisher: Telephone collection protocols, their uses and yields," in *Proc. Eurospeech*, 2003.

[3] O. Kimball, R. Iyer, and J. Makhoul, "From CTRAN to WordWave: More CTS for less," in *Proc. Rich Transcription Workshop*, 2003,

[4] S. Strassel, D. Miller, K. Walker, and C. Cieri, "Shared resources for robust speech-to-text technology," in *Proc. Eurospeech*, 2003.

[5] A. Stolcke, "Entropy-based pruning of backoff language models," in *Proc. DRAPA News Transcription and Understanding Workshop, Lansdowne*, 1998.

[6] I. Bulyko, M. Ostendorf, and A. Stolcke, "Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures," in *Proceedings of HLT*, 2003.

[7] X. Liu, M.J.F. Gales, K.C. Sim, and K. Yu, "Investigation of Acoustic Modelling techniques for LVCSR systems," in *submitted to ICASSP*, 2005.

[8] G. Evermann and P.C. Woodland, "Design of fast LVCSR systems," in *Proceedings ASRU*, 2003.