LARGE SCALE MMIE TRAINING FOR CONVERSATIONAL TELEPHONE SPEECH RECOGNITION

P.C. Woodland & D. Povey

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK.
Email: {pcw,dp10006}@eng.cam.ac.uk

ABSTRACT

This paper describes a lattice-based framework for maximum mutual information estimation (MMIE) of HMM parameters which has been used to train HMM systems for conversational telephone speech transcription using up to 265 hours of training data. These experiments represent the largest-scale application of discriminative training techniques for speech recognition of which the authors are aware, and have led to significant reductions in word error rate for both triphone and quinphone HMMs compared to our best models trained using maximum likelihood estimation. The use of MMIE training was a key contributer to the performance of the CU-HTK March 2000 Hub5 evaluation system.

1 INTRODUCTION

The model parameters in HMM based speech recognition systems are normally estimated using Maximum Likelihood Estimation (MLE). If certain conditions hold, including model correctness, then MLE can be shown to be optimal. However, when estimating the parameters of HMM-based speech recognisers, the true data source is not an HMM and therefore other training objective functions, in particular those that involve discriminative training, are of interest.

During MLE training, model parameters are adjusted to increase the likelihood of the word strings corresponding to the training utterances without taking account of the probability of other possible word strings. In contrast to MLE, discriminative training schemes, such as Maximum Mutual Information Estimation (MMIE) which is the focus of this paper, take account of possible competing word hypotheses and try and reduce the probability of incorrect hypotheses.

Discriminative schemes have been widely used in small vocabulary recognition tasks, where the relatively small number of competing hypotheses makes training viable. For large vocabulary tasks, especially on large datasets there are two main problems: generalisation to unseen data in order to increase test-set performance over MLE; and providing a viable computation framework to estimate confusable hypotheses and perform parameter estimation.

This paper is arranged as follows. First the details of the MMIE objective function are introduced. Then the lattice-based framework used for a compact encoding of alternative hypotheses is described along with the Extended Baum-Welch (EBW) algorithm for updating model parameters. Methods to enhance generalisation performance of MMIE trained systems are also discussed. Sets of experiments for evaluating the techniques on conversational telephone speech transcription are presented that show how MMIE training can be successfully applied over a range of training set sizes; the effect of methods to improve generalisation; and the interaction of MMIE with maximum-likelihood adaptation.

2 MMIE CRITERION

MMIE training was proposed in [1] as an alternative to MLE and maximises the mutual information between the training word sequences and the observation sequences. When the language model (LM) parameters are fixed during training (as they are in this paper and in almost all MMIE work in the literature), the MMIE criterion increases the *a posteriori* probability of the word sequence corresponding to the training data given the training data.

For R training observation sequences $\{\mathcal{O}_1, \ldots, \mathcal{O}_r, \ldots \mathcal{O}_R\}$ with corresponding transcriptions $\{w_r\}$, the MMIE objective function is given by

$$\mathcal{F}_{\mathrm{MMIE}}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(\mathcal{O}_r|\mathcal{M}_{w_r})P(w_r)}{\sum_{\hat{w}} p_\lambda(\mathcal{O}_r|\mathcal{M}_{\hat{w}})P(\hat{w})} \tag{1}$$

where $\mathcal{M}_w$ is the composite model corresponding to the word sequence w and $P(w)$ is the probability of this sequence as determined by the language model. The summation in the denominator of (1) is taken over all possible word sequences $\hat{w}$ allowed in the task and it can be replaced by

$$p_\lambda(\mathcal{O}_r|\mathcal{M}_{\mathrm{den}}) = \sum_{\hat{w}} p_\lambda(\mathcal{O}_r|\mathcal{M}_{\hat{w}})P(\hat{w}) \tag{2}$$

where $\mathcal{M}_{\mathrm{den}}$ encodes the full acoustic and language model used in recognition.

It should be noted that optimisation of (1) requires the maximisation of the numerator term $p_\lambda(\mathcal{O}_r|\mathcal{M}_{w_r})$, which is identical to the MLE objective function, while simultaneously minimising the denominator term $p_\lambda(\mathcal{O}|\mathcal{M}_{\mathrm{den}})$.

3 EXTENDED BAUM-WELCH ALGORITHM

The most effective method to optimise the MMIE objective function for large data and model sets is the Extended Baum-Welch (EBW) algorithm [3] as applied to Gaussian mixture HMMs [6].

The update equations for the mean of a particular dimension of the Gaussian for state j, mixture component m, μ_{jm} and the corresponding variance, σ^2_{jm} (assuming diagonal covariance matrices) can be re-estimated by

$$\hat{\mu}_{jm} = \frac{\left\{\theta^{\mathrm{num}}_{jm}(\mathcal{O}) - \theta^{\mathrm{den}}_{jm}(\mathcal{O})\right\} + D\mu_{jm}}{\left\{\gamma^{\mathrm{num}}_{jm} - \gamma^{\mathrm{den}}_{jm}\right\} + D} \tag{3}$$

$$\hat{\sigma}^2_{jm} = \frac{\left\{\theta^{\mathrm{num}}_{jm}(\mathcal{O}^2) - \theta^{\mathrm{den}}_{jm}(\mathcal{O}^2)\right\} + D(\sigma^2_{jm} + \mu^2_{jm})}{\left\{\gamma^{\mathrm{num}}_{jm} - \gamma^{\mathrm{den}}_{jm}\right\} + D} - \hat{\mu}^2_{jm} \tag{4}$$

In these equations, the $\theta_{j,m}(\mathcal{O})$ and $\theta_{j,m}(\mathcal{O}^2)$ are sums of data and squared data respectively, weighted by occupancy, for mixture component m of state j, and the Gaussian occupancies (summed over time) are γ_{jm}. The superscripts num and den refer to the model corresponding to the correct word sequence, and the recognition model for all word sequences, respectively.

It is important to have an appropriate value for D in the update equations, (3) and (4). If the value set is too large then training is very slow (but stable) and if it is too small the updates may not increase the objective function on each iteration. A useful lower bound on D is the value which ensures that all variances remain positive. Using a single global value of D can lead to very slow convergence, and in [9] a phone-specific value of D was used.

In preliminary experiments, it was found that the convergence speed could be further improved if D was set on a per-Gaussian level, i.e. a Gaussian specific D_{jm} was used. It was set at the maximum of i) twice the value necessary to ensure positive variance updates for all dimensions of the Gaussian; or ii) the denominator occupancy $\gamma_{jm}^{\mathrm{den}}$.

The mixture weight values were set using a novel approach described in [7]. The exact update rule for the mixture weights is not too important for the decision-tree tied-state mixture Gaussian HMMs used in the experiments reported here, since the Gaussian means and variances play a much larger role in discrimination.

4 IMPROVING MMIE GENERALISATION

An important issue in MMIE training is the ability to generalise to unseen test data. While MMIE training often greatly reduces training set error from an MLE baseline, the reduction in error rate on an independent test set is normally much less, i.e., compared to MLE, the generalisation performance is poorer. Furthermore, as with all statistical modelling approaches, the more complex the model, the poorer the generalisation. Since fairly complex models are needed to obtain optimal performance with MLE, it can be difficult to improve these with conventional MMIE training. We have considered two methods of improving generalisation that both increase the amount of confusable data processed during training: weaker language models and acoustic model scaling.

In [8] it was shown that improved test-set performance could be obtained using a unigram LM during MMIE training, even though a bigram or trigram was used during recognition. The aim is to provide more focus on the discrimination provided by the acoustic model by loosening the language model constraints. In this way, more confusable data is generated which improves generalisation. An unigram LM for MMIE training is investigated in this paper.

When combining the likelihoods from an HMM-based acoustic model and the LM it is usual to scale the LM log probability. This is necessary because, primarily due to invalid modelling assumptions, the HMM underestimates the probability of acoustic vector sequences. An alternative to LM scaling is to multiply the acoustic model log likelihood values by the inverse of the LM scale factor (acoustic model scaling). While this produces the same effect as language model scaling when considering only a single word sequence as for Viterbi decoding, when likelihoods from different sequences are added, such as in the forward-backward algorithm or for the denominator of (1), the effects of LM and acoustic model scaling are very different. If language model scaling is used, one particular state-sequence tends to dominate the likelihood at any point in time and hence dominates any sums using path likelihoods. However, if acoustic scaling is used, there will be several paths that have fairly similar likelihoods which make a non-negligible contribution to the summations. Therefore acoustic model scaling tends to increase the confusable data set in training by broadening the posterior distribution of state occupation $\gamma_{jm}^{\mathrm{den}}$ that is used in the EBW update equations. This increase in confusable data also leads to improved generalisation performance.

5 LATTICE-BASED MMIE TRAINING

The parameter re-estimation formulae presented in Section 3 require the generation of occupation and weighted data counts for both the numerator terms which rely on using the correct word sequence and the denominator terms which use the recognition model. The calculation of the denominator terms directly is computationally very expensive and so, in this work and as suggested in [9], word lattices are used to approximate the denominator model.

The first step is to generate word-level lattices, normally using an MLE-trained HMM system and a bigram LM appropriate for the training set. This step is normally performed just once and for the experiments in Section 6 the word lattices were generated in about 5x Real-Time (RT).[1]

The second step is to generate *phone-marked* lattices which label each word lattice arc with a phone/model sequence and the Viterbi segmentation points. These are are found from the word lattices and a particular HMM set, which may be different to the one used to generate the original word-level lattices. In our implementation, these phone marked lattices also encode the LM probabilities used in MMIE training which again may be different to the LM used to generate the original word-level lattices. This stage typically took about 2xRT to generate triphone-marked lattices for the experiments in Section 6, although the speed of this process could be considerably increased.

Given the phone-marked lattices for the numerator and denominator of each training audio segment, the lattice search used here performs a full forward-backward pass at the state-level constrained by the lattice and the statistics needed for the EBW updates accumulated. Pruning is performed by using the phone-marked lattice segmentation points extended by a short-period in each direction.[2] The search was also optimised as far as possible by combining redundantly repeated models which occur in the phone-marked lattice. Typically after compaction, the method requires about 1xRT per iteration for the experiments in Section 6.

6 MMIE EXPERIMENTS ON HUB5 DATA

This section describes a series of MMIE training experiments using the Cambridge University HTK (CU-HTK) system for the transcription of conversational telephone data from the Switchboard and Call Home English corpora ("Hub5" data). These experiments were performed in preparation for the NIST March 2000 Hub5 Evaluation. Details of the March 2000 CU-HTK Hub5 system can be found in [5].

The experiments investigated the effect of different training set and HMM set sizes and types; the use of acoustic likelihood scaling and unigram LMs in training and any possible interactions between MMIE training and maximum likelihood linear regression-based adaptation.

---

[1] All run times are measured on an Intel Pentium III running at 550MHz.

[2] Typically 50ms at both the start and end of each phone.

6.1 Basic CU-HTK Hub5 System

The CU-HTK Hub5 system is a continuous mixture density, tied-state cross-word context-dependent HMM system based on the HTK HMM Toolkit. The full system operates in multiple passes, using more complex acoustic and language models and unsupervised adaptation in later passes.

Incoming speech is parameterised into cepstral coefficients and their first and second derivatives to form a 39 dimensional vector every 10ms. Cepstral mean and variance normalisation and vocal tract length normalisation is performed for each conversation side in both training and test.

The HMMs are constructed using decision-tree based state-clustering and both triphone and quinphone models can be used. All experiments here used gender independent HMM sets. The pronunciation dictionary used in the experiments discussed below was for either a 27k vocabulary (as used in [4]) or a 54k vocabulary and the core of this dictionary is based on the LIMSI 1993 WSJ lexicon. The system uses word-based N-gram LMs estimated from an interpolation of Hub5 acoustic training transcriptions and Broadcast News texts. In the experiments reported here, trigram LMs are used unless otherwise stated.

6.2 Experiments with 18 Hours Training

Initially we investigated MMIE training using the 18 hour BBN-defined Minitrain corpus with an HMM set using 3088 speech states and 12 Gaussian/state HMMs, which were our best MLE trained models. Lattices were generated on the training set using a bigram LM. The bigram 1-best hypotheses had a 24.6% word error rate (WER) and a Lattice WER (LWER) of 6.2%.

MMIE Iteration	%WER	
	Acoustic Scaling	LM Scaling
0 (MLE)	50.6	50.6
1	50.2	51.0
2	49.9	51.3
3	50.5	51.4
4	50.9	–

Table 1: 18 hour experiments with 12 mixture component models (eval97sub): comparison of acoustic model and LM scaling.

The Minitrain 12 Gaussian/state results given in Table 1 compare acoustic and language model scaling for several iterations of MMIE training on the eval97sub test set (a subset of the 1997 Hub5 evaluation). It can be seen that acoustic scaling helps avoid over-training and the best WER is after 2 iterations. The training set lattices regenerated after a single MMIE iteration gave a WER of 16.8% and a LWER of 3.2%, showing that the technique is very effective in reducing training set error. However, it was found that these regenerated lattices were no better to use in subsequent training iterations and so all further work used just the initially generated word lattices.

The advantage of MMIE training for the 12 Gaussian per state system is small and so the same system with 6 Gaussians/state was trained. The results in Table 2 and again show the best performance after two MMIE iterations. Furthermore the gain over the MLE system is 1.7% absolute if a bigram LM is used and 1.9% absolute if a unigram LM is used: the 6 Gaussian per state MMIE-trained HMM set now slightly outperforms the 12 Gaussian system. Furthermore it can be seen that using a weakened LM (unigram) improves performance a little.

MMIE Iteration	%WER	
	Lattice Bigram	Lattice Unigram
0 (MLE)	51.5	51.5
1	50.0	49.7
2	49.8	49.6
3	50.1	50.0
4	50.8	–

Table 2: 18 hour experiments with 6 mixture component models (eval97sub): comparison of lattice LMs.

6.3 Experiments with 68 Hours Training

The effect of extending the training set to the 68 hour h5train00sub set [5] was investigated next using an HMM system with 6165 speech states and 12 Gaussians/state. Tests were performed on both the eval97sub and the 1998 evaluation set (eval98). In this case the phone-marked denominator lattices had a LWER of 7.4%.

MMIE Iteration	%WER	
	eval97sub	eval98
0 (MLE)	46.0	46.5
1	43.8	45.0
2	43.7	44.6
3	44.1	44.7

Table 3: Word error rates on eval97sub and eval98 using h5train00sub training.

The results in Table 3 show that again the peak improvement comes after two iterations, but there is an even larger reduction in WER: 2.3% absolute on eval97sub and 1.9% absolute on eval98. The word error rate for the 1-best hypothesis from the original bigram word lattices measured on 10% of the training data was 27.4%. The MMIE models obtained after two iterations on the same portion of training data gave an error rate of 21.2%, so again MMIE provided a very sizeable reduction in training set error.

6.4 Triphone Experiments with 265 Hours Training

The good performance on smaller training sets led us to investigate MMIE training using all the available Hub5 data: the 265 hour h5train00 set. The h5train00 set contains 267,611 segments and numerator and denominator word level lattices were created for each trained segment, and from these, phone-marked lattices were generated. The HMMs used here had 6165 speech states and 16 Gaussians/state.

MMIE Iteration	%WER	
	eval97sub	eval98
0 (MLE)	44.4	45.6
1	42.4	43.7
1 (3xCHE)	42.0	43.5
2	41.8	42.9
2 (3xCHE)	41.9	42.7

Table 4: Word error rates when using h5train00 training with and without CHE data weighting (3xCHE).

We also experimented with data-weighting with this setup during MMIE training. The rationale for this is that while the test data sets contain equal amounts of Switchboard and CHE data, the training set is not balanced. Therefore we gave a 3x higher weighting to CHE data during training. The results of these experiments on both the eval97sub and eval98 test sets are shown in Table 4. It can be seen that there is an improvement in WER of 2.6% absolute on eval97sub and 2.7% on eval98.

Data weighting gives a further small improvement, although interestingly, data weighting for MLE reduces the WER by 0.7% absolute on eval97sub. It might be concluded that the extra weight placed on poorly recognised data by MMIE training relative to MLE reduces the need for the data weighting technique.

6.5 Quinphone Model Training

Since the CU-HTK Hub5 system can use quinphone models, we investigated MMIE quinphone training using h5train00. The decision tree state clustering process for quinphones includes questions regarding ± 2 phone context and word-boundaries. The baseline quinphone system uses 9640 speech states and 16 Gaussians/state.

The quinphone MMIE training used triphone-generated word lattices, but, since the phone-marked lattices were re-generated for the quinphone models, it was necessary to further prune the word-lattices. The results of MMIE trained quinphones on the eval97sub set are shown in Table 5. Note that these experiments, unlike all previous ones reported here, include pronunciation probabilities.

MMIE Iteration	%WER eval97sub
0 (MLE)	42.0
1	40.4
2	39.9
3	40.1

Table 5: Quinphone MMIE results on eval97sub. CHE data weighting used for MLE baseline.

As with the MMIE training runs discussed above, the largest WER reduction (2.1% absolute) comes after two iterations of training. The reductions in error rate are similar to those seen for triphone models when CHE data weighting is used even though there was extra pruning required for the phone-marked lattices and there were rather more HMM parameters to estimate.

6.6 Interaction with MLLR

All the above results used models that were not adapted to the particular conversation side using maximum likelihood linear regression (MLLR) [2]. To measure MLLR adaptation performance, MMIE and MLE models (with data weighting) were used in a full-decode of the test data, i.e. not rescoring lattices, with a 4-gram language model. The output from this first pass was used to estimate a global speech MLLR (block-diagonal mean and diagonal variance) transform using the output from the respective non-adapted pass was used for adaptation supervision. The adapted models were then used for a second full-decode pass.

The results in Table 6 show that the MMIE models are 2.1% absolute better than the MLE models without MLLR, and 2.2% better with MLLR. In this case, MLLR seems to work just as well with MMIE trained models: a relatively small number of parameters are being estimated with MLLR and these global transforms keep the Gaussians in the same "configuration" as optimised by MMIE.

Adaptation	% WER eval98	
	MLE	MMIE
None	44.6	42.5
MLLR	42.1	39.9

Table 6: Effect of MLLR on MLE and MMIE trained models.

7 CONCLUSIONS

This paper has discussed the use of discriminative training for large vocabulary HMM-based speech recognition for a training set size and level of task difficulty not previously attempted. It has been shown that 2-3% absolute reductions in word error rates can be obtained for the transcription of conversational telephone speech. The use of HMMs trained using MMIE was the most significant addition to the March 2000 CU-HTK evaluation system.

Acknowledgements

This work is in part supported by a grant from GCHQ. Dan Povey holds a studentship from the Schiff Foundation.

References

1. L.R. Bahl, P.F. Brown, P.V. de Souza & R.L. Mercer (1986). Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition, *Proc. ICASSP'86*, pp. 49–52, Tokyo.
2. M.J.F. Gales & P.C. Woodland (1996). Mean and Variance Adaptation Within the MLLR Framework. *Computer Speech & Language*, Vol. 10, pp. 249–264.
3. P.S. Gopalakrishnan, D. Kanevsky, A. Nádas & D. Nahamoo (1991). An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems. *IEEE Trans. Information Theory*, Vol. 37, pp. 107–113.
4. T. Hain, P.C. Woodland, T.R. Niesler & E.W.D. Whittaker (1999). The 1998 HTK System for Transcription of Conversational Telephone Speech. *Proc. ICASSP'99*, pp. 57–60, Phoenix.
5. T. Hain, P.C. Woodland, G. Evermann & D. Povey (2000). The CU-HTK March 2000 Hub5E Transcription System. *Proc. Speech Transcription Workshop*, College Park.
6. Y. Normandin (1991). *Hidden Markov Models, Maximum Mutual Information Estimation and the Speech Recognition Problem.* Ph.D. Thesis, McGill University, Montreal.
7. D. Povey & P.C. Woodland (1999). *An Investigation of Frame Discrimination for Continuous Speech Recognition.* Technical Report CUED/F-INFENG/TR.332, Cambridge University Engineering Dept.
8. R. Schlüter, B. Müller, F. Wessel & H. Ney (1999). Interdependence of Language Models and Discriminative Training. *Proc. IEEE ASRU Workshop*, pp. 119–122, Keystone, Colorado.
9. V. Valtchev, J.J. Odell, P.C. Woodland & S.J. Young (1997). MMIE Training of Large Vocabulary Speech Recognition Systems. *Speech Communication*, Vol. 22, pp. 303–314.