IMPROVED DISCRIMINATIVE TRAINING TECHNIQUES FOR LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

D. Povey & P.C. Woodland

Cambridge University Engineering Dept, Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {dp10006,pcw}@eng.cam.ac.uk

ABSTRACT

This paper investigates the use of discriminative training techniques for large vocabulary speech recogntion with training datasets up to 265 hours. Techniques for improving lattice-based Maximum Mutual Information Estimation (MMIE) training are described and compared to Frame Discrimination (FD). An objective function which is an interpolation of MMIE and standard Maximum Likelihood Estimation (MLE) is also discussed. Experimental results on both the Switchboard and North American Business News tasks show that MMIE training can yield significant performance improvements over standard MLE even for the most complex speech recognition problems with very large training sets.

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) [1, 5, 9] which is the main 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 computationally tractable. 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, followed by a description of the training scheme used for optimisation and methods to enhance generalisation performance of MMIE trained systems. MMIE systems are compared both to systems trained with the frame discrimination (FD) technique [6] and to training using an interpolation of the MLE and MMIE criteria. The training schemes are evaluated using both the Switchboard/ Call Home English (CHE) corpus and North American Business News (NAB) data.

2. MMIE OBJECTIVE FUNCTION

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.

For R training observation sequences $\{\mathcal{O}_1, \dots, \mathcal{O}_r, \dots \mathcal{O}_R\}$ 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 $p_\lambda(\mathcal{O}_r|\mathcal{M}_{\mathrm{den}})$.

3. MMI OPTIMISATION

The most effective method to optimise the MMIE objective function for large data and model sets is the Extended Baum-Welch (EBW) algorithm [2] as applied to Gaussian mixture HMMs [5].

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), are as follows:

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

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

In these equations, $\theta_{j,m}(\mathcal{O})$ and $\theta_{j,m}(\mathcal{O}^2)$ are weighted sums of data and squared data respectively, for mixture component m of state j, where the weighting is by posterior probability of Gaussian occupation at each time. The summed Gaussian posterior probabilities 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 is too large, training is very slow (but stable), but if it is too small the updates may not increase the objective function at 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 therefore in [9] a phone-specific value of D was used. We have found that the convergence speed can be further improved if D is set on a per-Gaussian level, i.e. a Gaussian specific D_{jm} used. In this work, D_{jm} was set at the maximum of i) twice the value necessary to ensure positive variance updates for all dimensions of the Gaussian; and ii) a further constant E times the denominator occupancy $\gamma_{j,m}^{\mathrm{den}}$ for that Gaussian. For experiments reported here, E=2 was used.

The mixture weight values were set using a novel approach described in [7], and informal experiments have shown that normally it results in a faster increase in the overall MMIE objective function than the use of the standard updating formula used in e.g. [9]. However, the update rule for the mixture weights is less 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 discriminative 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, both of which 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. A unigram LM for MMIE training is used 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. 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 for 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 9 the word lattices were generated in about 5x Real-Time (RT) for the Switchboard experiments and 1.5xRT for the NAB experiments[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 Switchboard experiments and 0.5xRT for the NAB lattices, 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 start and end times for each phone. The search was also optimised by combining redundantly repeated models which occur in the phone-marked lattice with the same start and end times. Typically after compaction, the method requires about 0.4xRT per iteration for the Switchboard experiments and 0.1xRT per iteration for the NAB experiments reported in Section 9.

6. FRAME DISCRIMINATION

In FD training [4, 6], the MMIE denominator is replaced by a single state HMM which is the weighted sum of all states in the original HMM system. Optimisation of this function is performed in the same manner as for MMIE.

The weights assigned to each state output distribution in the system are derived from alignments of the training data generated during Maximum Likelihood training. This very general denominator model leads to good generalisation at the expense of poorer training set performance. Furthermore, exact computation of the denominator requires calculation of all the Gaussians in the system for each frame, and for large HMM systems it is necessary to

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

approximate the denominator using just the most likely Gaussians. In [6] an efficient algorithm was developed for this purpose.

Acoustic model scaling can also be used with FD. For the results reported for Switchboard a scaling factor of 0.5 was used, since without acoustic scaling FD training was found to offer no reduction in word error rate (WER) over MLE. However for comparison with the results in [6], the FD results on the NAB corpus use no acoustic scaling.

7. INTERPOLATED OBJECTIVE FUNCTIONS

The use of the MMIE objective function is rather prone to cause *over-training*, so that after a few iterations the error rate on independent test data starts to increase. Therefore MMIE training is normally stopped before over-training occurs. In contrast, MLE is much less prone to over-training, and therefore objective functions which are a combination of MMIE and MLE could be of interest.

The idea of forming an interpolated MMIE/MLE objective function is related to the idea of the H-criterion in [3]. Given the re-estimation equations (3) and (4), it is straightforward to implement an objective function which is an interpolation of the MMIE and MLE criteria. The denominator part of the EBW equations simply needs to be scaled, since the numerator alone represents the required statistics for the MLE criterion. For instance, to implement an objective function which is $0.9\mathcal{F}_{\text{MMIE}}+0.1\mathcal{F}_{\text{MLE}}$, the denominator of the EBW equations is scaled by a factor of 0.9.

8. EXPERIMENTAL SETUP

The following sections describe the experimental framework for both the Switchboard experiments and those on NAB data. In both cases the input data consists of PLP coefficients derived from a mel-scale filter bank (MF-PLP), with 13 coefficients including c_0 and their first and second-order differentials. The HMMs used were gender independent cross-word triphones built using decision-tree state clustering. Conventional MLE was used to initialise the HMMs prior to discriminative training. For both sets of experiments, word lattices for MMIE training were created using a bigram language model, while unigram probabilities were actually applied to these lattices for MMIE training.

Recognition experiments used lattice rescoring of word lattices derived using MLE HMMs. The pronunciation dictionaries used in training and test were originally based on the 1993 LIMSI WSJ lexicon, but have been considerably extended and modified.

8.1. NAB System

The NAB experiments used HMMs trained on the SI-284 Wall Street Journal database (66 hours of data) and used per-utterance cepstral mean normalisation. The HMMs used in the NAB experiments were the same as those used in [6] to test FD training. These HMMs have 6399 speech states and versions of these models with 1,2,4 and 12 mixture components per state were used.

The NAB experiments used the 1994 DARPA Hub1 development and evaluation test sets, denoted csrnab1_dt and csrnab1_et respectively, and used lattice rescoring of the same 65k word vocabulary trigram lattices previously used in [6] for strict comparability with those results.

8.2. Switchboard System

For the Switchboard experiments, we used two training sets comprising of a total of 265 hours of data taken from the Switchboard1 and Call Home English corpora. Further details of this training corpus, denoted h5train00, are given in [10]. Most experiments were performed with a 68 hour subset, denoted h5train00sub. The data had cepstral mean and variance normalisation applied on a conversation side basis, along with vocal tract length normalisation. The HMMs used had 6165 clustered speech states and 12 Gaussians per state for h5train00sub training and 16 Gaussians per state when using h5train00.

Two test sets were used for the experiments: eval97sub, a subset of the 1997 Hub5 evaluation set, containing 10 conversation sides of Switchboard2 (Swb2) data and 10 of CHE; and the 1998 Hub5 evaluation data set, eval98, containing 40 sides of SWB2 and 40 CHE sides (in total about 3 hours of data). Recognition used a 27k word vocabulary with a language model formed by an interpolation of Switchboard and Broadcast News LMs.

9. EXPERIMENTAL RESULTS

9.1. NAB results

Table 1 gives recognition results for MMIE training for the NAB corpus. The results for HMM sets with a number of mixture components are included after 4 iterations of MMIE. The table also contains the results using MLE and the FD results from [6].

#Mix	csrnab1_dt_h1			csrnab1_et_h1		
Comp	MLE	FD	MMIE	MLE	FD	MMIE
1	13.64	11.95	11.36	15.64	14.32	13.16
2	11.84	10.58	10.12	13.19	12.04	11.31
4	10.67	9.77	9.42	11.25	10.84	10.59
12	9.30	8.99	8.80	9.96	9.85	9.40
12.ns	9.30	—	9.23	9.96	—	9.61

Table 1. % WER using MMIE and FD training for various NAB model sets. All tests use unigram LMs and acoustic likelihood scaling apart from 12.ns which used language model scaling and a bigram LM in MMIE training

The relative improvement due to MMIE, averaged over both test sets, varies from 16.3% in the single Gaussian system to 5.5% in the 12 mixture component system. However, importantly, there is still a worthwhile improvement over the best MLE system. Also, unlike the case discussed in [6], the current implementation of MMIE outperforms the results from FD in all cases. Hence the generalisation performance of the system is much better than the MMIE experiments in [9] which used a bigram language model, no acoustic model scaling and different alignment and update procedures. The effect of not using the acoustic model scaling with a bigram language model is given in the line marked 12.ns in Table 1 where the performance is, on average, the same as FD.

9.2. Switchboard results

The effect of applying MMIE training to Switchboard is given in Table 2 for several iterations of MMIE updating. Furthermore, previous experiments in [10] showed that generalisation performance is a greater issue with conversational telephone Switchboard data than with the much cleaner read newspaper texts in the NAB corpus: if acoustic scaling and a unigram language model aren't used on Switchboard then no performance improvements result from MMIE training of the most complex models.

Iteration	h5train00sub		h5train00	
Number	eval97sub	eval98	eval97sub	eval98
0	46.0	46.6	44.4	45.6
1	44.4	45.4	42.6	44.0
2	43.7	44.7	41.9	42.9
3	43.9	44.4	41.6	42.7
4	43.9	44.3	41.4	42.2

Table 2. % WER from several iterations of MMIE training on the h5train00 and h5train00sub data sets.

On average, a 2.3% absolute reduction in WER is obtained using the 68-hour training setup and 3.2% absolute using the 265-hour setup. However if training is continued there is some evidence of over-training. Therefore results are reported only up to the 4th training iteration. Note also that these results are somewhat better than those we previously obtained for the 265-hour setup [10].

The use of an interpolated objective function was then investigated using the 68 hour h5train00sub setup with various proportions of the MMIE objective function, and training was continued until the eighth iteration. The recognition results from these models are given in Table 3, which gives error rates for several values of the proportion of the MMIE function x (i.e, $x\mathcal{F}_{\rm MMIE} + (1 - x)\mathcal{F}_{\rm MLE}$).

	x = fraction MMIE				
	1.0	0.9	0.8	0.7	0.5
eval97sub	44.2	44.0	44.1	43.6	43.9
eval98	44.3	44.0	44.0	44.2	44.9

Table 3. % WER for an interpolated (MMIE and MLE) objective function for different values of the MMIE interpolation weight. The h5train00sub training set was used.

The WER results for a proportion of 0.7 MMIE and 0.3 MLE are slightly better than the best results in Table 2 for the h5train00sub setup. Unsurprisingly, the HMM model parameters for these interpolated models are closer to the original MLE models than using pure MMIE. It was found that the average parameter difference from the MLE model set varied roughly linearly with the proportion of MMIE used in the overall objective function. As well as giving similar performance to MMIE, a model set that is closer to the MLE set may have advantages, for instance, with procedures that assume maximum likelihood parameter training such as some adaptation and confidence estimation techniques.

Finally the effect of FD using the h5train00sub database was investigated. On eval97sub, after 4 iterations, an error rate of 44.8% was obtained, compared to 43.9% for MMIE and 46.0% from MLE. The computational cost per iteration of FD was roughly 1.5xRT for this data and therefore is about 3 times as expensive as the iterations of MMIE. However MMIE first requires lattice creation but this can be done once for several MMIE experiments and training iterations. For comparison, MLE training runs in about 0.07xRT per iteration.

10. CONCLUSIONS

This paper has shown that significant improvements over standard maximum likelihood training can be obtained using discriminative training techniques for large vocabulary tasks with very large data sets. For good performance, it is important to take steps to improve MMIE generalisation using acoustic likelihood scaling and weakened language models. While greater performance improvements occur for simpler acoustic models, the techniques described are able to provide reductions in word error rate over the best-performing MLE models. MMIE training is shown to be more effective than the use of frame discrimination. The use of an interpolated objective function may lead to small further improvements over MMIE.

11. ACKNOWLEDGEMENTS

This work is in part supported by a grant from GCHQ and by the EU Framework 5 project Coretex. Dan Povey holds a studentship from the Schiff Foundation.

12. 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] P.S. Gopalakrishnan, D. Kanevsky, A. Nadas & D. Nahamoo (1991). An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems. *IEEE Trans. on Information Theory*, Vol. 37, pp 107-113.

[3] P.S. Gopalakrishnan, D. Kanevsky, A. Nadas, D. Nahamoo & M.A. Picheney (1988). Decoder Selection Based on Cross-Entropies. *Proc. ICASSP'88*, Vol. 1, pp. 20-23.

[4] S. Kapadia (1998). *Discriminative Training of Hidden Markov Models*. Ph.D. thesis, Cambridge University Engineering Dept.

[5] Y. Normandin (1991). *Hidden Markov Models, Maximum Mutual Information Estimation and the Speech Recognition Problem.* Ph.D. thesis, Dept. of Elect. Eng., McGill University, Montreal.

[6] D. Povey & P.C. Woodland (1999). Frame Discrimination Training of HMMs for Large Vocabulary Speech Recognition. *Proc. ICASSP'99*, pp. 333–336, Phoenix.

[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.

[10] P.C. Woodland & D. Povey (2000). Large Scale Discriminative Training for Speech Recognition. *Proc. ISCA ITRW ASR2000*, pp. 7–16, Paris.