LATTICE-BASED DISCRIMINATIVE TRAINING FOR LARGE VOCABULARY SPEECH RECOGNITION

V. Valtchev, J.J. Odell †, P.C. Woodland, & S.J. Young

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, England.

†Entropic Cambridge Research Laboratory,
Sheraton House, Castle Park, Cambridge, CB3 0AX, England.

ABSTRACT

This paper describes a framework for optimising the parameters of a continuous density HMM-based large vocabulary recognition system using a Maximum Mutual Information Estimation (MMIE) criterion. To limit the computational complexity arising from the need to find confusable speech segments in the large search space of alternative utterance hypotheses, word lattices generated from the training data are used. Experiments are presented on the Wall Street Journal database using up to 66 hours of training data. These show that lattices combined with an improved estimation algorithm makes MMIE training practicable even for very complex recognition systems and large training sets. Furthermore, experimental results show that MMIE training can yield useful increases in recognition accuracy.

1. INTRODUCTION

Previous research has shown that the accuracy of a speech recognition system trained using Maximum Likelihood Estimation (MLE) can often be improved further using discriminative training. In particular, Maximum Mutual Information Estimation (MMIE) has been studied in the context of small vocabulary speech tasks and substantial gains in performance have been reported [1, 2].

Discriminative optimisation of HMM parameters using MMIE is much more complex than the corresponding MLE case. Firstly, given the complete data set, there is no closed form for the optimal parameters as there is in Baum-Welch re-estimation. Instead, some form of gradient-based optimisation must be used. Thus, whilst an MLE system can typically be trained in a few iterations, MMIE training may require considerably more. Secondly, discriminative training requires confusion data representative of the recognition errors made by the speech recognition device. Even in a small vocabulary task, the gathering of statistics about mismatched segments of speech results in a dramatic increase in computational requirements compared to the corresponding MLE case.

The focus of the work reported here is to find methods of improving the recognition accuracy of large vocabulary continuous speech recognition (LVCSR) systems which typically have several million parameters and require very large training databases. Thus, the application of discriminative training techniques to LVCSR systems is computationally extremely challenging.

A recent trend in the design of LVCSR systems has been the inclusion of facilities to generate lattices encoding multiple recognition hypotheses. Currently used for system development purposes or multi-pass recognition, these lattices provide a compact encoding of confusion data and therefore offer a route towards making MMIE training of such systems practicable.

This paper explores the use of lattices for the discriminative training of LVCSR systems. The basic framework is described and then a number of experiments using the HTK tied-state LVCSR system [6] and the Wall Street Journal database are presented.

2. MMI ESTIMATION OF HMM PARAMETERS

MMIE training attempts to increase the *a posteriori* probability of the model sequence corresponding to the training data given the training data. For R training observations $\{\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_r, \ldots \mathcal{O}_R\}$ the MMIE objective function is given by

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

and it is usually assumed that the denominator in equation (1) can be replaced by

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

where $\mathcal{M}_{\hat{w}}$ is the model corresponding to the word sequence $\hat{w}$ and $\mathcal{M}_{gen}$ is a model constructed such that for all paths in every $\mathcal{M}_{\hat{w}}$ there is a corresponding path of equal probability in $\mathcal{M}_{gen}$. Typically, $\mathcal{M}_{gen}$ is the model used during recognition. Thus, MMIE training can be interpreted as a two stage optimisation process. The first stage is equivalent to performing MLE training such that the HMM parameters are adapted to increase the numerator term $P_\lambda(\mathcal{O}_r|\mathcal{M}_r)$. In the second stage, the HMM parameters are changed in the opposite direction in order to minimise the denominator term $P_\lambda(\mathcal{O}|\mathcal{M}_{gen})$. The second step dominates the computation and this will depend on the size of the vocabulary, the grammar and any contextual constraints. In many practical situations, for example where cross-word context dependent models are used in conjunction with a long span language model, the construction of a complete model for $\mathcal{M}_{gen}$ is intractable.

A continuous density HMM system can be optimised according to the above MMIE objective function by using the following equations to re-estimate the means $\mu_{j,m}$ and variances $\sigma^2_{j,m}$ for each state j and mixture component m [2].

$$\hat{\mu}_{j,m} = \frac{\left\{\theta_{j,m}(\mathcal{O}) - \theta^{gen}_{j,m}(\mathcal{O})\right\} + D\mu_{j,m}}{\left\{\psi_{j,m} - \psi^{gen}_{j,m}\right\} + D} \tag{3}$$

$$\hat{\sigma}^2_{j,m} = \qquad (4)$$
$$\frac{\left\{\theta_{j,m}(\mathcal{O}^2) - \theta^{gen}_{j,m}(\mathcal{O}^2)\right\} + D(\sigma^2_{j,m} + \mu^2_{j,m})}{\left\{\psi_{j,m} - \psi^{gen}_{j,m}\right\} + D} - \hat{\mu}^2_{j,m}$$

where $\theta_{j,m}(x)$ represents the sum of all x weighted by the probability of occupying component m of state j and $\psi_{j,m}$ represents the corresponding occupation counts.

Similarly, the mixture weight parameters $c_{j,m}$ can be re-estimated according to

$$\hat{c}_{j,m} = \frac{c_{j,m}\left\{\frac{\partial}{\partial c_{j,m}}\mathcal{F}(\lambda) + C\right\}}{\sum_{\hat{m}} c_{j,\hat{m}}\left\{\frac{\partial}{\partial c_{j,\hat{m}}}\mathcal{F}(\lambda) + C\right\}} \qquad (5)$$

To remove emphasis from small-valued parameters [2], the derivatives of the objective function are approximated by the following expression

$$\frac{\partial}{\partial c_{j,m}}\mathcal{F}(\lambda) \approx \frac{\gamma_{j,m}}{\sum_{\hat{m}}\gamma_{j,\hat{m}}} - \frac{\gamma^{gen}_{j,m}}{\sum_{\hat{m}}\gamma^{gen}_{j,\hat{m}}} \qquad (6)$$

where $\gamma_{j,m}$ is the occupancy count of mixture component m in state j.

The constant D is set to be just large enough to ensure that all variances remain positive. The constant C is chosen such that all parameter derivatives are positive. Experimentally, these selection criteria have been shown to give relatively smooth and fast convergence [1, 2].

3. WORD LATTICES

A word lattice forms a compact representation of many different sentence hypotheses and hence provides an efficient representation of the confusion data needed for discriminative training [3]. In the HTK system [6], a lattice consists of a set of nodes that correspond to particular instants in time, and arcs connecting these nodes to represent possible word hypotheses. Associated with each arc is an acoustic score (log likelihood) and a language model score.

Lattices are generated as a by-product of the recognition process. The HTK LVCSR system uses a time-synchronous one-pass decoder that is implemented using a dynamically built tree-structured recognition network. This approach allows the integration of cross-word context-dependent acoustic models and an N-gram language model directly within the search [4]. Once these lattices are constructed, they can be used as a word graph to constrain the search space in further recognition passes. Assuming that the lattice coverage does not change during parameter re-estimation, this use of lattices as a constraining word graph forms the basis of the proposed MMIE training algorithm which is as follows:

1. Generate a pair of *numerator* and *denominator* lattices for each utterance in the training data, these correspond to $\mathcal{M}_r$ and $\mathcal{M}_{gen}$, respectively. The numerator lattice is produced by aligning the acoustic data against a network of HMMs built according to the "correct" transcription. The denominator lattice corresponds to running an unconstrained recognition pass. In both cases an appropriate N-gram language model is used.
2. For each training utterance, the numerator or denominator lattice is loaded into the recogniser and reduced to a word graph. Recognition is performed using the current HMM set and the language model scores from the word graph. A new output lattice is then produced containing the original language model scores and new acoustic scores. For each node in the lattice and unique spanning word $w_{i,j}$, the forward ($\bar{\alpha}$) and the backward ($\bar{\beta}$) lattice probabilities are computed. The forward probabilities are given by
$$\bar{\alpha}_j = \sum_i \bar{\alpha}_i P_{acoust}(w_{ij}) P_{lang}(w_{ij}) \qquad (7)$$
and the backward probabilities are computed in a similar fashion starting from the end of the lattice. For each pair of nodes i and j, the corresponding $\bar{\alpha}_i$ and $\bar{\beta}_j$ are propagated into the sequence of model instances corresponding to word $w_{i,j}$, and statistics are accumulated.
3. The two sets of statistics accumulated by performing step 2 separately for the numerator and denominator of the MMIE objective function are combined together and new parameter estimates calculated according to equations (3), (4) and (5).

4. SYSTEM DEVELOPMENT

This section describes the practical development of the discriminative training framework. The development was carried out in two separate stages - generation of training set lattices and optimisation of HMM parameters.

4.1. Lattice Generation

The lattices were generated using the HTK LVCSR system. The system uses state-clustered, cross-word mixture Gaussian triphonic acoustic models and a back-off bigram language model.

Each frame of speech is represented by a 39 dimensional feature vector that consists of 12 mel frequency cepstral coefficients, normalised log energy and the first and second differentials of these values.

The state clustering algorithm uses decision trees built for every monophone HMM state to determine equivalence classes between sets of triphone contexts. This is followed by the application of an iterative mixture splitting and re-training sequence which allows the optimal match between system complexity and available training data to be found.

For the lattice generation on the training data, a 65k word list was created by adding the words occurring in the training set to our standard WSJ recognition lexicon [6]. A corresponding bigram back-off language model was then constructed to accommodate the SI-284 training set which contains utterances with both verbalised and non-verbalised punctuation. The language model (`train_bg65k`) contained 4.2 million bigrams estimated from the `nab94` text corpus of 227 million words.

Two sets of lattices were generated, each consisting of *numerator* and *denominator* subsets. Table 1 gives a comparison of the HMM systems used to generate each set. Table 2 gives an indication of the quality of the lattices in terms of word/sentence error rate and lattice density. The lattice density figure is the average number of lattice arcs (representing words) per spoken word. The lattice sentence error rate relates to whether a path corresponding to the

System	States	#Mix. comps.	Data
HMM-0	3948	2	SI-84 (WSJ0)
HMM-1	6399	12	SI-284 (WSJ0+1)

Table 1. HMM systems used for lattice generation.

correct sentence transcription exists in the lattice. The lattice word error rate is a lower bound on the word error rate from rescoring the lattice.

System	Set	Density	%SER	%WER
HMM-0/1	num	1.7	0.0	0.0
HMM-0	den	50.7	21.6	1.9
HMM-1	den	14.9	14.8	1.2

Table 2. Lattice densities and % lattice sentence/word error rates for the HMM-0/1 lattice sets.

Both HMM-0 and HMM-1 are gender independent state-clustered cross-word triphone systems using the 1993 LIMSI WSJ Lexicon and phone set. The HMM-0 system was trained on the SI-84 (WSJ0) data set (7,176 utterances). A relatively wide pruning beam was used which resulted in an average lattice density figure of 50.7. The HMM-1 system was trained on the SI-284 (WSJ0+1) data set (36,441 utterances). Details of this system and its performance on various WSJ test sets is given in [6]. Lattices were generated for the full SI-284 training set. To speed up development time, a tight pruning beam was used resulting in an average lattice density figure of 14.9.

Finally, some of the denominator lattices in the HMM-0/1 sets were found not to contain the correct transcription of the utterance. To solve this problem, the corresponding numerator and denominator lattices were merged together to form a new set of denominator lattices used for MMIE training.

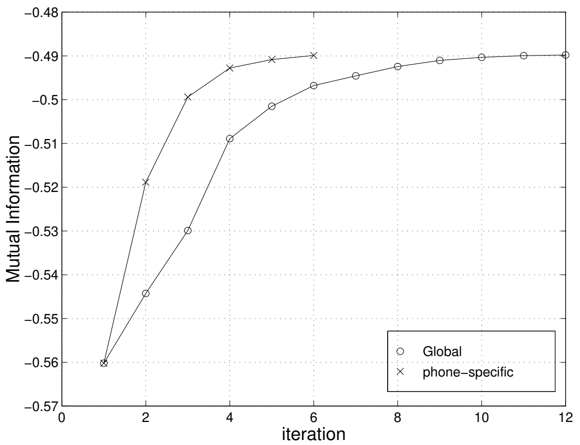

Figure 1. Change in the value of the MMIE objective function for global and per-phone D constants using HMM-0 with 2 mixture components and 400 training utterances from SI-84.

4.2. MMIE optimisation

Preliminary experiments were performed in order to establish and tune the convergence properties of the MMI re-estimation algorithm. The optimisation was focused on the mean/variance parameters of the Gaussian distributions. The speed of convergence is directly related to the value of the constant D in the re-estimation formulae (equations 3 and 4). These tuning experiments were performed using the HMM-0 system with 2 mixture components per state on a subset of 400 utterances from the SI-84 (WSJ0) training set.

Similarly to [1], in the first experiment a global constant was used such that all re-estimated variance parameters were positive. In the second experiment phone-specific constants were used (47 different constants in total) such that the variance parameters for all triphones of each phone were positive. The value of the objective function in the two cases is plotted in Figure 1. It is clear from the plot that the use of phone-specific constants improves the convergence rate of the algorithm by a factor of two.

5. EXPERIMENTS

All recognition experiments were performed on the 1995 American English evaluation test set (`si_et`) selected for the European SQALE project [5] which consist of 200 utterances from a total of 20 different talkers. All results are scored using the official SQALE word-map file.

5.1. HMM-0 experiments

The preliminary evaluation of the proposed MMIE training method was carried out using the family of HMM-0 state-clustered cross-word context dependent triphone systems with 2, 4 and 8 mixture components per state. The models were originally trained for 4 iterations using the MLE criterion on the SI-84 (WSJ0) data set. In each case, this was followed by 10 iterations of MMIE training for mean, variance and mixture weight parameters. In general, it was found that the best performance figures were obtained after the fourth iteration of MMIE training, after which, the performance started to degrade. Recognition test were run using the ARPA November 1993 20k word list and associated bigram language model estimated from 37 million words of WSJ text data supplied by MIT Lincoln Labs. The test set results from all experiments were produced by re-scoring recognition lattices originally computed using the baseline HMM-0 system with 8 mixture components per state.

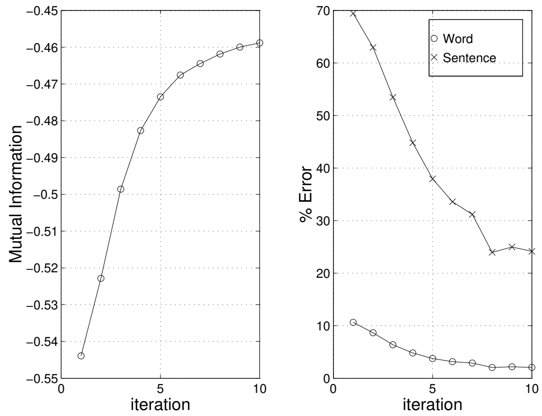

Figure 2. Objective function value and lattice recognition performance on the SI-84 training set using the HMM-0 system with 2 mixture components per state.

The results in Table 3 shows the recognition performance of the systems on the training data (SI-84). The plot in

Figure 2 also shows the typical change in the value of the objective function at each iteration together with the recognition performance of the system in terms of word/sentence error rates.

Table 4 shows the recognition performance of the various systems on the `si_et` test set in increasing order of complexity. Despite possible gains in performance obtained by rescoring lattices from a more sophisticated system the results should be indicative of the relative improvements in performance contributed by the MMIE training scheme.

The results show that MMIE has provided a worthwhile improvement in the performance of all systems with the resulting performance of the 2 mixture MMIE trained system matching that of the original 8 mixture MLE trained system. The 4 mixture MMIE system provides the best overall word error rate of 17.4%. Unfortunately the improvement in the performance of the 8 mixture system after MMIE training was found to be small. Indeed, this is not too surprising in view of the limited amount of training data used.

#Mix. comps.	MLE		MMIE	
	%WER	%SER	%WER	%SER
2	10.6	69.4	2.1	24.2
4	8.0	60.7	1.4	17.1
8	5.4	49.4	1.1	13.5

Table 3. Recognition results on the SI-84 training set using the HMM-0 systems with different number of mixture components per state and the train_bg65k LM. WER/SER denote word/sentence error rate respectively.

#Mix. comps.	MLE		MMIE	
	%WER	%SER	%WER	%SER
2	20.6	83.5	18.0	78.5
4	18.9	82.0	17.4	79.0
8	18.0	79.5	17.5	80.0

Table 4. Recognition results on the si_et test set using the HMM-0 systems and the 1993 20k ARPA bigram LM.

5.2. HMM-1 experiments

The HMM-1 system in these experiments used the combined WSJ0+1 corpora resulting in 66 hours of acoustic training data and a recognition vocabulary of 65k words with bigram, trigram and fourgram language models estimated from the `nab94` text corpus of 227 million words (`test_65k` LMs). The use of a larger recognition vocabulary reduces the OOV rate from 1.46% to 0.39% for the `si_et` test set. This combined with the substantial amount of acoustic and language model training data results in an improvement in baseline recognition from 18.0% to 12.6% word error rate for the best performing MLE trained systems using bigram LMs.

The HMM-1 system was optimised for four iterations of MMIE training. The first row in Table 5 shows the performance of the system on the training set using the `train_bg65k` bigram LM. Consistent with previous MMIE results, the word error rate is dramatically reduced from 8.1% to 3.5%. The second, third and fourth rows in Table 5 show the performance of the system on the `si_et` test set with 65k bigram, trigram and fourgram language models respectively. In all cases, the MMIE training has resulted in improved recognition performance.

Data set	LM	MLE		MMIE	
		%WER	%SER	%WER	%SER
`SI-284`	bg	8.1	58.8	3.5	34.1
`si_et`	bg	12.6	77.0	11.9	73.5
`si_et`	tg	9.0	60.0	8.2	59.5
`si_et`	fg	7.9	56.0	7.4	55.0

Table 5. Recognition results on the SI-284 training set (train_bg65k LM) and si_et test set (test_65k LMs) using the HMM-1 system.

6. CONCLUSIONS

This paper has described an implementation of MMIE discriminative training based on the use of lattices to compactly encode and compute the required confusion data. It has been demonstrated that this approach makes it feasible to apply MMIE training to very large HMM-based recognition systems. Furthermore, the re-estimation formulae used previously for small systems give good convergence on large systems provided that the learning rate constants for mean and variance parameters are set on a per phone basis.

Experimental results using the WSJ0 and WSJ0+1 sets of the Wall Street Journal training database show that the proposed method is very effective in reducing the word error rate on the training set. Typically, results on unseen test data show a reduction in word error rate of 5-10% following MMIE training.

ACKNOWLEDGMENTS

This work is in part supported by an EPSRC/MOD research grant GR/J10204. Additional computational resources were provided by the ARPA CAIP computing facility.

REFERENCES

[1] S. Kapadia, V. Valtchev, and S.J. Young. MMI Training for Continuous Phoneme Recognition on the TIMIT Database. In *Proc. ICASSP'93*, volume 2, pages 491–494, Minneapolis, April 1993. IEEE.

[2] Y. Normandin. *Hidden Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Problem.* PhD thesis, Department of Electrical Engineering, McGill University, Montreal, March 1991.

[3] Y. Normandin, R. Lacouture, and R. Cardin. MMIE Training for Large Vocabulary Continuous Speech Recognition. In *Proc. ICSLP'94*, volume 3, pages 1367–1370, Yokohama, September 1994.

[4] J.J. Odell, V. Valtchev, P.C. Woodland, and S.J. Young. A One Pass Decoder Design For Large Vocabulary Recognition. In *Proc. ARPA Human Language Technology Workshop*, pages 405–410. Morgan Kaufmann, March 1994.

[5] D. Pye, P.C. Woodland, and S.J. Young. Large Vocabulary Multilingual Speech Recognition using HTK. In *Proc. EUROSPEECH'95*, volume 1, pages 181–184, Madrid, September 1995.

[6] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev, and S.J. Young. The 1994 HTK Large Vocabulary Speech Recognition System. In *Proc. ICASSP'95*, volume 1, pages 73–76, Detroit, May 1995. IEEE.