MINIMUM PHONE ERROR AND I-SMOOTHING FOR IMPROVED DISCRIMINATIVE TRAINING

D. Povey & P.C. Woodland

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

ABSTRACT

In this paper we introduce the Minimum Phone Error (MPE) and Minimum Word Error (MWE) criteria for the discriminative training of HMM systems. The MPE/MWE criteria are smoothed approximations to the phone or word error rate respectively. We also discuss I-smoothing which is a novel technique for smoothing discriminative training criteria using statistics for maximum likelihood estimation (MLE). Experiments have been performed on the Switchboard/Call Home corpora of telephone conversations with up to 265 hours of training data. It is shown that for the maximum mutual information estimation (MMIE) criterion, I-smoothing reduces the word error rate (WER) by 0.4% absolute over the MMIE baseline. The combination of MPE and I-smoothing gives an improvement of 1% over MMIE and a total reduction in WER of 4.8% absolute over the original MLE system.

1. INTRODUCTION

Model parameters in HMM-based speech recognition systems are normally estimated using Maximum Likelihood Estimation (MLE). However, since the conditions for MLE optimality, including model correctness, do not hold, other optimisation criteria are of interest. Over the years, several *discriminative training* criteria, including Maximum Mutual Information Estimation (MMIE) [1, 5] and Minimum Classification Error (MCE) [4, 6], have been successfully applied to small vocabulary speech recognition tasks.

Until recently it was believed that discriminative training techniques are not effective in reducing the word error rate (WER) for the most difficult large vocabulary tasks using HMM systems with a very large number of parameters. The key issues are a viable computational framework which allows incorrect word hypotheses to be efficiently processed and good generalisation to test data. It was shown in [10] that the computation can be made viable by using a lattice-based framework along with the Extended Baum-Welch (EBW) algorithm [3, 5] for MMIE parameter estimation. Generalisation can be improved by using acoustic scaling to increase the effective amount of confusable data [7, 11] and a weak unigram language model (LM) during training [9]. It was demonstrated [7, 11] that these techniques together yield reduced WER over the best MLE systems for large vocabulary tasks.

While we have previously focused on MMIE, this paper proposes techniques that, like MCE, minimise an estimate of the training set errors. For some small tasks, it has been reported [8, 9] that MCE outperforms MMIE. However, we know of no experiments using MCE for large vocabulary speech recognition. Indeed since MCE targets the sentence error rate, the implicit weight assigned to each frame of data has an undesirable dependence on the training data segmentation into utterances.

As an alternative to MCE we have developed the Minimum Word Error (MWE) objective function. MWE maximises the expected word accuracy and can be easily computed in a lattice framework. We have also developed the Minimum Phone Error (MPE) criterion which uses the same approach at the phone level.

The paper also discusses *I-smoothing* which applies smoothing between the discriminative and MLE estimates for a parameter in a way such that the degree of smoothing depends on the amount of data available. While this is beneficial to MMIE, it is essential to make MWE/MPE outperform standard MLE training.

The paper first introduces the MWE/MPE objective functions and discusses their optimisation in a lattice context. The use of I-smoothing is then described. Experiments on the transcription of telephone conversations are then presented which show the effectiveness of the current methods.

2. MWE/MPE OBJECTIVE FUNCTIONS

This section describes the various objective functions used in this paper. For R training observation sequences $\{\mathcal{O}_1,\ldots,\mathcal{O}_r,\ldots\mathcal{O}_R\}$ with corresponding transcriptions $\{s_r\}$, the MMIE objective function for HMM parameter set λ, including the effect of scaling the acoustic and LM probabilities[1] can be written

$$\mathcal{F}_{\mathrm{MMIE}}(\lambda)=\sum_{r=1}^{R}\log\frac{p_\lambda\left(\mathcal{O}_r|\mathcal{M}_{s_r}\right)^\kappa P(s_r)^\kappa}{\sum_s p_\lambda\left(\mathcal{O}_r|\mathcal{M}_s\right)^\kappa P(s)^\kappa} \qquad (1)$$

where $\mathcal{M}_s$ is the composite model corresponding to the word sequence s and $P(s)$ 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 allowed in the task. Hence MMIE maximises the posterior probability of the correct sentences. The denominator in (1) can be approximated by a word lattice of alternative sentence hypotheses.

The MWE criterion is defined as

$$\mathcal{F}_{\mathrm{MWE}}(\lambda)=\sum_{r}^{R}\frac{\sum_s p_\lambda(\mathcal{O}_r|s)^\kappa P(s)^\kappa \mathrm{RawAccuracy}(s)}{\sum_s p_\lambda(\mathcal{O}_r|s)^\kappa P(s)^\kappa}$$

where $\mathrm{RawAccuracy}(s)$ is a measure of the number of words accurately transcribed in hypothesis s. Hence, for each training

Dan Povey was supported by the Schiff foundation. Grants from GCHQ & the EU project Coretex have also contributed to this work. It has benefited from an SUR award from IBM.

[1] It is assumed that the LM probabilities $P(s)$ have already been "scaled" (raised to the power) by the normal LM scale factor $1/\kappa$ and hence further scaling by κ takes them back to their original values.

0-7803-7402-9/02/$17.00 ©2002 IEEE

utterance, the MWE criterion gives a weighted average over all s of the RawAccuracy(s). Ideally this is the the metric used to calculate WER, i.e the number of correct words in s minus the number of insertions. Then when $\kappa \rightarrow \infty$, maximising the MWE criterion becomes equivalent to minimising the word error rate. A key issue is how to define RawAccuracy(s) so that it avoids dynamic programming for each hypothesis and can be efficiently implemented in a lattice-based framework.

As well as the MWE criterion, we have also investigated the Minimum Phone Error (MPE) criterion, which uses the same approach as MWE but estimates errors at the phone level. Either context-independent (CI) or context dependent (CD) phone labels can be used, leading to either MPE-CI or MPE-CD.

3. OPTIMISATION OF DISCRIMINATIVE CRITERIA

This section describes the optimisation strategy for both MMIE and MWE. The approach allows the MMIE EBW re-estimation formulae to be used for MWE given suitable "occupancies".

3.1. Optimisation of the MMIE objective function

To gather the statistics needed for the EBW algorithm, for each training utterance the lattice corresponding to either the numerator (num) or the denominator (den) of (1) are used to compute the posterior probabilities of occupation of mixture component m of state j at time t, e.g. $\gamma_{j,m}^{\text{num}}(t)$. These are then used to gather Gaussian occupancies and weighted sums of the data and squared data $\theta_{j,m}^{\text{num}}(\mathcal{O})$ and $\theta_{j,m}^{\text{num}}(\mathcal{O}^2)$ respectively.

The statistics gathering process uses the *exact-match* forward-backward procedure [11]. This uses the phone boundary times from the lattice and applies the scale factor κ to lattice phone arc log likelihoods. The forward-backward procedure is performed first between the start and end times of phone arcs q, leading to within-arc posterior probabilities given the arc, which in the numerator case are denoted $\gamma_{j,m,q}^{\text{num}}(t)$. The within-arc process also generates a likelihood $p(q)$ for that arc. These arc likelihoods together with probabilities arising from the language model are scaled by the factor κ and used in a forward-backward pass at the lattice-node level to estimate the arc posterior probability, γ_q^{num} (i.e. the probability of traversing that arc). The overall occupancies needed for the EBW formulae are then gathered according to formulae such as $\gamma_{j,m}^{\text{num}} = \sum_{q=1}^{Q} \sum_{t=s_q}^{e_q} \gamma_{j,m,q}^{\text{num}}(t)\gamma_q$.

Once the statistics have been accumulated over all training files, they are used to update the Gaussian parameters as follows [11]:

$$\hat{\mu}_{j,m} = \frac{\{\theta_{j,m}^{\text{num}}(\mathcal{O}) - \theta_{j,m}^{\text{den}}(\mathcal{O})\} + D\mu_{j,m}}{\{\gamma_{j,m}^{\text{num}} - \gamma_{j,m}^{\text{den}}\} + D},$$

$$\hat{\sigma}_{j,m}^2 = \frac{\{\theta_{j,m}^{\text{num}}(\mathcal{O}^2) - \theta_{j,m}^{\text{den}}(\mathcal{O}^2)\} + D(\sigma_{j,m}^2 + \mu_{j,m}^2)}{\{\gamma_{j,m}^{\text{num}} - \gamma_{j,m}^{\text{den}}\} + D} - \hat{\mu}_{j,m}^2$$

The constant D is set on a per-Gaussian level to the greater of: i) twice the smallest value needed to ensure positive variances, or ii) $\gamma_{j,m}^{\text{den}}$ times a further constant E, which is generally set to 1 or 2.

Updated mixture weights $\hat{c}_{j,m}$ are calculated by maximising the following auxiliary function:

$$\sum_{m=1}^{M} \gamma_{j,m}^{\text{num}} \log \hat{c}_{j,m} - \frac{\gamma_{j,m}^{\text{den}}}{c_{j,m}} \hat{c}_{j,m},$$

subject to the sum-to-one constraint; a similar method is used for each row of the transition matrices.

3.2. Optimisation of the MWE objective function

In MMIE training, the difference $\gamma_q^{\text{num}} - \gamma_q^{\text{den}}$ between the two arc posterior probabilities is equal to to $\frac{1}{\kappa}\frac{\partial \mathcal{F}_{\text{MMIE}}}{\partial \log p(q)}$, where $p(q)$ is the likelihood of the speech data from the beginning to the end of the phone arc. The approach taken for optimisation of the MWE objective function is to calculate for each phone arc the value of

$$\gamma_q^{\text{MWE}} = \frac{1}{\kappa} \frac{\partial \mathcal{F}_{\text{MWE}}}{\partial \log p(q)}$$

which will be denoted the "MWE arc occupancy". If positive, then we set $\gamma_q^{\text{num}} = \gamma_q^{\text{MWE}}$, $\gamma_q^{\text{den}} = 0$ and add to the numerator EBW statistics only; if negative, then set $\gamma_q^{\text{num}} = 0$, $\gamma_q^{\text{den}} = -\gamma_q^{\text{MWE}}$, and add to the denominator statistics. The statistics thus obtained are then used in the EBW parameter update equations just as they would be for MMIE training. MWE requires a forward-backward pass over just the denominator lattice[2] rather than both the numerator and denominator although time-alignment information from the numerator is included in the MWE occupancy computation (see Sec. 4). Each arc in the denominator lattice will contribute either to the numerator or denominator statistics, depending on the sign of γ_q^{MWE}. In the next section, it will be explained how to calculate the "MWE occupancies" γ_q^{MWE} in a lattice framework.

4. CALCULATING MWE ARC OCCUPANCIES

MWE arc occupancies are easily computed if the RawAccuracy(s) can be expressed as a sum of terms each corresponding to a word w regardless of the context, i.e. we require that RawAccuracy(s) = $\sum_{w \in s}$ WordAcc(w), where ideally we would have:

$$\text{WordAcc}(w) = \left\{ \begin{array}{l} 1 \text{ if correct word} \\ 0 \text{ if substitution} \\ -1 \text{ if insertion} \end{array} \right\}.$$

Since the computation of the above expression requires dynamic programming, the value used here is as follows. A word z is found in the reference transcript which overlaps in time with hypothesis word w, then if the proportion of the length of z which is overlapped is denoted e, set

$$\text{WordAcc}(w) = \left\{ \begin{array}{l} -1 + 2e \text{ if same word} \\ -1 + e \text{ if different word} \end{array} \right\}. \quad (2)$$

The word z is chosen so as to make WordAcc(w) as large as possible. The expressions in (2) represent tradeoffs between an insertion and a correct word or substitution respectively, and are a solution to the problem that a single reference word might be used more than once by a hypothesis sentence. In our implementation the reference word z is chosen from a lattice encoding alternate alignments of the correct sentence.

Differentiation of the MWE objective function leads to an expression for γ_q^{MWE} as follows:

$$\gamma_q^{\text{MWE}} = \gamma_q(c(q) - c_{\text{avg}}),$$

[2] Note that, as for MMIE training, the correct sentence hypothesis is added to the denominator lattice if not already present.

where γ_q is the arc occupancy as derived from a forward backward pass over the arcs, $c(q)$ is the average value of RawAccuracy(s) for sentences s containing arc q (weighted by the κ-scaled log likelihood of those sentences), and c_{avg} is the weighted average RawAccuracy(s) for all sentences in the lattice, which is the same as the MWE criterion for the utterance.

The value of $c(w)$ may be efficiently calculated by another lattice forward-backward pass. Since the WordAcc(w) in (2) is defined for words and the forward-backward algorithm will work at the phone level, let us define PhoneAcc(q) to be, in the case of MWE, WordAcc(w) if q is first phone of w, and zero otherwise. In the case of MPE, PhoneAcc(w) would be calculated directly from an equation of the form in (2). Then, if α_q and β_q are the forward and backward likelihoods used to calculate normal arc posterior probabilities, let

$$\begin{aligned} \alpha'_q &= \frac{\sum_{r \text{ preceding } q} \alpha'_r \alpha_r t^{\kappa}_{rq}}{\sum_{r \text{ preceding } q} \alpha_r t^{\kappa}_{rq}} + \text{PhoneAcc}(q) \\ \beta'_q &= \frac{\sum_{r \text{ following } q} t^{\kappa}_{qr} p(r)^{\kappa} \beta_r (\beta'_r + \text{PhoneAcc}(r))}{\sum_{r \text{ following } q} t^{\kappa}_{qr} p(r)^{\kappa} \beta_r} \\ c(q) &= \alpha'_q + \beta'_q. \end{aligned}$$

where t_{qr} are lattice transition probabilities derived from the language model and κ is the likelihood scale.

5. I-SMOOTHING

The H-criterion [2] uses a fixed interpolation between the MLE (H=0) and MMIE (H=1) objective functions. For the large training sets we have investigated, we haven't found it reduces WER, although it is useful as a technique to make MMIE training converge without over-training [11].

I-smoothing is a way of applying an interpolation between MLE and a discriminative objective function in a way which depends on the amount of data available for each Gaussian. In the context of MMIE, I-smoothing simply means increasing the number of data points $\gamma^{\text{num}}_{j,m}$ assigned to Gaussian j, m by τ while keeping the average data values and average squared data values the same; in the context of MPE training, it involves adding τ points of the MLE occupancies (as obtained from the alignment of the correct transcriptions) to the numerator occupancies $\gamma^{\text{num}}_{j,m}$, $\theta^{\text{num}}_{j,m}(\mathcal{O})$ and $\theta^{\text{num}}_{j,m}(\mathcal{O}^2)$ used in MPE training. In the MPE case, this would be done as follows:

$$\begin{aligned} \gamma'^{\text{num}}_{j,m} &= \gamma^{\text{num}}_{j,m} + \tau \\ \theta'^{\text{num}}_{j,m}(\mathcal{O}) &= \theta^{\text{num}}_{j,m}(\mathcal{O}) + \frac{\tau}{\gamma^{\text{mle}}_{j,m}} \theta^{\text{mle}}_{j,m}(\mathcal{O}) \\ \theta'^{\text{num}}_{j,m}(\mathcal{O}^2) &= \theta^{\text{num}}_{j,m}(\mathcal{O}^2) + \frac{\tau}{\gamma^{\text{mle}}_{j,m}} \theta^{\text{mle}}_{j,m}(\mathcal{O}^2) \end{aligned}$$

where the superscript mle indicates occupancies as would be obtained by alignment of the correct transcriptions. In the MMIE case, I-smoothing is applied by increasing all of $\gamma^{\text{num}}_{j,m}$, $\theta^{\text{num}}_{j,m}(\mathcal{O})$ and $\theta^{\text{num}}_{j,m}(\mathcal{O}^2)$ by a factor $1 + \frac{\tau}{\gamma^{\text{num}}_{j,m}}$. In both cases the EBW parameter update equations are then applied using the altered counts.

A technique very similar in effect to I-smoothing but not involving arbitrary constants has been developed based on a Maximum A Posteriori principle. The technique gives a justification of the I-smoothing process, and the particular range of τ found in practice to be effective.

6. EXPERIMENTAL SETUP

To evaluate the discriminative training techniques experiments have been performed on the transcription of "Hub5" from the Switchboard and Call Home English (CHE) corpora. The basic setup is the same as used for MMIE experiments reported in [7, 11].

The input speech 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. Word lattices for discriminative training were created using a bigram LM, while unigram probabilities were actually applied to these lattices during training. In all experiments, the scale value κ is set to the inverse of the standard recognition LM scale factor. The discriminative training schemes were generally tested after 8 iterations of updating unless otherwise shown.

We used two training sets comprising of a total of 265 hours of data taken from the Switchboard1 and CHE corpora. Further details of this training corpus, denoted h5train00, are given in [11]. 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 with 12 Gaussians per state for h5train00sub training and 16 Gaussians per state when using h5train00.

Recognition experiments used 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. The 1998 Hub5 evaluation data set, eval98, was used for testing. This contains 40 sides of Switchboard2 and 40 CHE sides (in total about 3 hours of data). Recognition used a 27k word vocabulary with a trigram language model formed by an interpolation of Switchboard and Broadcast News LMs.

We also report recognition results on a subset of the training data. This uses about 2 hours of training data that was randomly selected from the training corpora. The training results use either a full (fast) single pass decode using a bigram LM, rescoring the training word lattices using a bigram LM or rescoring the actual unigram LM lattices used in discriminative training.

7. RESULTS

Table 1 shows both the training and test WERs for training on either a) the 68 hour or b) the full 265 hour training set for standard MMIE, MMIE with I-smoothing and MPE. For larger amounts of data, MPE gives the greatest reduction in training set WER on the unigram lattices on which the system is trained. However, it does not give as large a reduction in training set WER as MMIE when tested with a bigram language model. It should be noted that the full-decode and lattice bigram decoding results are similar.

I-smoothing improves MMIE test-set performance (by about 0.5% absolute) at the cost of training set accuracy i.e. it gives improved generalisation. The use of the MPE objective function further improves test-set accuracy: with the full training set it gives a 1% reduction in WER over standard MMIE. It should be noted that the value of τ at which the best results are obtained for MPE (e.g, $\tau = 50$) represents at least as much smoothing as the, say,

Training Type (training iteration)	WER Training Subset			WER Test
	Full bg	Lat bg	Lat ug	eval98
MLE baseline	26.3	26.0	41.8	46.6
MMIE E=2,τ=0 (4)	18.6	19.4	30.1	44.3
MMIE E=1,τ=200 (6)	19.7	20.3	32.2	43.8
MPE E=2,τ=50 (8)	20.6	20.7	27.9	43.1

(a)

Training Type (training iteration)	WER Training Subset			WER Test
	Full bg	Lat bg	Lat ug	eval98
MLE baseline	30.1	29.8	47.2	45.6
MMIE E=2,τ=0 (8)	23.2	23.7	37.7	41.8
MMIE E=1,τ=200 (8)	22.2	23.0	35.8	41.4
MPE E=2,τ=100 (8)	23.9	23.9	34.4	40.8

(b)

Table 1. Training & test WERs for MMIE (τ=0), I-smoothed MMIE and MPE for (a) 68 hour and (b) 265 hour training set.

$\tau = 200$ in the MMIE case; the figures are not comparable because occupancies in MWE tend to be considerably less than one.

Training Type	Train Subset WER		MPE Train	Test WER
	Full bg	Lat ug	Criterion	eval98
MLE baseline	26.3	41.8	0.66	46.6
MPE E=2,τ=0	25.5	28.5	0.80	50.7
MPE E=2,τ=25	20.0	26.2	0.81	43.1
MPE E=2,τ=50	20.6	27.9	0.79	43.1
MPE E=2,τ=100	21.6	29.9	0.77	43.3

Table 2. MPE with varying amounts of I-smoothing for 68 hour training set.

Table 2 shows the effect of varying the amount of I-smoothing on MPE. Without I-smoothing, MPE causes a degradation in test set performance by the 8th iteration. Training results from rescoring the training lattices and criterion improvement show that I-smoothing works by improving generalisation rather than criterion optimisation, and that good optimisation of the MPE criterion is not sufficient to obtain good test set results. Unsmoothed MPE seems to be less robust than MMIE since it does not generalise well to the use of a bigram LM on the training set, unlike MMIE.

Training Type	Training Subset WER		Test WER
	Full bg	Lat ug	eval98
MLE baseline	26.3	41.8	46.6
MWE E=2,τ=25	20.2	25.9	43.3
MPE $E = 2$,τ=50	20.6	27.9	43.1
MPE $E = 2$,τ=100	21.6	29.9	43.3
MPE-CD,E=2,τ=100	20.7	28.5	43.4

Table 3. MPE compared with MWE and context-dependent MPE using 68 hour training set.

Table 3 compares MPE with MWE and context-dependent MPE (MPE-CD). Although MPE-CD improves training set WERs, there is little difference in test-set performance. As expected, MWE produces a greater improvement in training set word error rate since word error rate is being more directly optimised. However, this is not matched by better performance on the test set.

8. CONCLUSIONS

Two new discriminative training criteria, Minimum Phone Error and Minimum Word Error, have been presented and a lattice-based implementation has been described. Both of these methods directly optimise a smoothed approximation of the training set errors. The focus on training errors, rather than posterior probability of the correct utterance as in MMIE, tends to place more weight on training data that is close to decision boundaries and might be corrected by small changes in the HMM parameter values. A technique called I-smoothing has been described which improves the generalisation of discriminatively trained HMMs and seems to be essential for MPE/MWE. I-smoothed MPE is currently our most effective discriminative training technique with a reduction in WER 4.8% absolute over MLE when trained on 265h of Switchboard/CHE data and a 1% absolute lower WER than our previous best MMIE result without I-smoothing.

9. 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. Nádas, D. Nahamoo & M.A. Picheney (1988). Decoder Selection Based on Cross-Entropies, *Proc. ICASSP'88*, Vol. 1, pp. 20-23, New York.

[3] 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.

[4] E. McDermott E & S. Katagiri (1994). Prototype-based Minimum Classification Error/Generalised Probabilistic Descent Training for Various Speech Units. *Computer Speech & Language*, Vol. 8, pp. 351-368.

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

[6] B.H. Juang, W. Chou & C.H. Lee (1997) Minimum Classification Error Rate Methods for Speech Recognition. *IEEE Trans SAP*, Vol. 5, pp. 266-277 .

[7] D. Povey & P.C. Woodland (2001). Improved Discriminative Training Techniques for Large Vocabulary Continuous Speech Recognition, *Proc. ICASSP'01*, Vol. 1, pp. 45-48, Salt Lake City.

[8] W. Reichl & G. Ruske (1995). Discriminative Training for Continuous Speech Recognition, *Proc. Eurospeech'95*, Vol. 1, pp. 537-540, Madrid.

[9] R. Schlüter (2000). *Investigations on Discriminative Training*. PhD. thesis, Aachen university.

[10] 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.

[11] P.C. Woodland & D. Povey (2002). Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition. *Computer Speech & Language*, Vol. 16, pp. 25-47.