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 confusible 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 \( \{O_1, \ldots, O_r, \ldots, O_R\} \) with corresponding transcriptions \( \{s_r\} \), the MIE objective function for HMM parameter set \( \lambda \), including the effect of scaling the acoustic and LM probabilities\(^1\) can be written

\[
F_{\text{MMIE}}(\lambda) = \sum_{r=1}^{R} \log \frac{p_{\lambda}(O_r | M_{s_r}) n s_r}{\sum_{s} p_{\lambda}(O_r | M_s)^n P(s)^n} \tag{1}
\]

where \( M_s \) 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

\[
F_{\text{MWE}}(\lambda) = \sum_{r} \frac{p_{\lambda}(O_r | \beta)^n P(s)^n \text{RawAccuracy}(s)}{\sum_{s} p_{\lambda}(O_r | \beta)^n P(s)^n} \tag{2}
\]

where \( \text{RawAccuracy}(s) \) is a measure of the number of words accurately transcribed in hypothesis \( s \). Hence, for each training task, the MMIE target is to maximise the probability of the correct sentence, whereas the MWE target is to maximise the expected word accuracy.

\( ^{1} \) 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 \( \kappa \) takes them back to their original values.
utterance, the MWE criterion gives a weighted average over all
s of the RawAccuray(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 \( k \to \infty \), 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 ap-
proach as MWE but estimates errors at the phone level. Either
twice the smallest value needed to ensure positive variances, or ii)
implemented in a lattice-based framework.

A key issue is how to define Minimum Phone Error (MPE) criterion, which uses the same ap-
proach as MWE but estimates errors at the phone level. Either
twice the smallest value needed to ensure positive variances, or ii)
implemented in a lattice-based framework.

\[ \gamma_q = \frac{1}{k} \frac{\partial \text{MWE}}{\partial \log p(q)} \]

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 numer-
ator 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 \( \gamma_q^{\text{MWE}} \). In the next section, it will be explained how to
calculate the “MWE occupancies” \( \gamma_q^{\text{MWE}} \) 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) =
\( \sum_{w} \text{WordAcc}(w) \), where ideally we would have:

\[ \text{WordAcc}(w) = \begin{cases} 1 & \text{if correct word} \\ 0 & \text{if substitution} \\ -1 & \text{if insertion} \end{cases} \]

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 over-
lapped is denoted \( \epsilon \), set

\[ \text{WordAcc}(w) = \begin{cases} -1 + 2\epsilon & \text{if same word} \\ -1 + \epsilon & \text{if different word} \end{cases} \]  \( 2\)

The word z is chosen so as to make WordAcc(w) as large as possible.
The expressions in (2) represent tradeoffs between an inser-
tion 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 ex-
pression for \( \gamma_q^{\text{MWE}} \) as follows:

\[ \gamma_q^{\text{MWE}} = \gamma_q(c(q) - c_{\text{avg}}) \]

Note that, as for MMIE training, the correct sentence hypothesis is
added to the denominator lattice if not already present.
where $\gamma_s$ 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 $\kappa$-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 $\text{WordAcc}(w)$ in (2) is defined for words and the forward-backward algorithm will work at the phone level, let us define $\text{PhoneAcc}(q)$ to be, in the case of MWE, $\text{WordAcc}(w)$ if $q$ is first phone of $w$, and zero otherwise. In the case of MPE, $\text{PhoneAcc}(w)$ would be calculated directly from an equation of the form in (2). Then, if $\alpha_q$ and $\beta_d$ are the forward and backward likelihoods used to calculate normal arc posterior probabilities, let

$$ q^* = \frac{\sum_{r, p} \text{prev.} \cdot \kappa \cdot \alpha_{r,p} \cdot \beta_d}{\sum_{r, p} \text{prev.} \cdot \kappa \cdot \alpha_{r,p} \cdot \beta_d + \text{PhoneAcc}(q)} $$

$$ \beta_q^* = \frac{\sum_r \text{following} \cdot t_{r,p} \cdot \alpha_r \cdot \beta_d (\gamma_s + \text{PhoneAcc}(r))}{\sum_r \text{following} \cdot t_{r,p} \cdot \alpha_r \cdot \beta_d} $$

$$ c(q) = \alpha_q^* + \beta_d^* $$

where $t_{r,p}$ are lattice transition probabilities derived from the language model and $\kappa$ 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 the HMM, 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 $\gamma_{ij,m}$ assigned to Gaussian $j$, $m$ by $\tau$ while keeping the average data values and average squared data values the same: in the context of MPE training, it involves adding $\tau$ points of the MLE occupancies (as obtained from the alignment of the correct transcriptions) to the numerator occupancies $\theta_{j,m}^{\text{num}}$ of $\theta_{j,m}^\text{num}$ and $\theta_{j,m}^\text{num}(O)$ used in MPE training. In the MPE case, this would be done as follows:

$$ \gamma_{j,m}^{\text{num}} = \gamma_{j,m}^{\text{num}} + \tau $$

$$ \theta_{j,m}^{\text{num}}(O) = \theta_{j,m}^{\text{num}}(O) + \frac{\tau}{\gamma_{j,m}^{\text{num}}} \theta_{j,m}^{\text{den}}(O) $$

$$ \theta_{j,m}^{\text{num}}(O^2) = \theta_{j,m}^{\text{num}}(O^2) + \frac{\tau}{\gamma_{j,m}^{\text{num}}} \theta_{j,m}^{\text{den}}(O^2) $$

where the superscript $\text{num}$ indicates occupancies as would be obtained by alignment of the correct transcriptions. In the MMIE case, I-smoothing is applied by increasing all of $\gamma_{j,m}^{\text{num}}$, $\theta_{j,m}^{\text{num}}(O)$ and $\theta_{j,m}^{\text{num}}(O^2)$ by a factor $1 + \frac{\tau}{\gamma_{j,m}^{\text{num}}}$. 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 $\tau$ 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 $\kappa$ 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 subset is made up of 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 $\tau$ at which the best results are obtained for MPE (e.g., $\tau = 50$) represents at least as much smoothing as the, say,
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.

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