RIVAL TRAINING: EFFICIENT USE OF DATA IN DISCRIMINATIVE TRAINING

Carsten Meyer  Georg Rose
Philips Research Laboratories
Weißhausstr. 2, D-52066 Aachen, Germany
e-mail: {Carsten.Meyer, Georg.Rose}@philips.com

ABSTRACT

We evaluate a simple extension of the corrective training algorithm for reestimation of the acoustic parameters, using — in addition to misrecognized sentences — also a selection of correctly recognized sentences for discrimination. Our approach (called “rival training”) is implementationally much less expensive than lattice-based discriminative training methods, since we apply a “hard” threshold criterion to select a subset of sentences for which a single competitor is used for discrimination. Still, significant performance gains are obtained compared to maximum likelihood and corrective training even for triphone models with 61 densities per mixture (on a digit string and a large vocabulary isolated word recognition task). Furthermore, the hard selection scheme may be used to accelerate the training process due to faster convergence and by restricting the training process to a fixed subset of training utterances.

1. INTRODUCTION

In a number of applications, discriminative training has been shown to substantially improve recognition performance as compared to maximum likelihood (ML) training [1, 2, 3, 7, 10]. The simplest algorithm, corrective training (CT) [2], can be implemented within the Viterbi formalism. Using acoustic models with a large number of mixture densities, however, only small performance gains are obtained if the training error rate is already low, since only misrecognized sentences are used for discrimination. This might not provide enough material for a robust reestimation of the large set of acoustic parameters involved in such models. Other discriminative training methods like Maximum Mutual Information (MMI) [1] and Minimum Classification Error (MCE) [5], using essentially every sentence, for a reestimation of the acoustic parameters, lead to better performance (see e.g. [9]), but are implementationally much more expensive since they operate on word lattices and thus require a weighting of observation vectors.

The aim of our work is to extend the set of sentences used for discrimination without requiring a weighting of observation vectors, thus enabling a simple implementation within the Viterbi formalism. By appropriately selecting the competitors a significant performance gain is obtained even for the most complex acoustic models.

The paper is organized as follows: In section 2 we sketch the basic idea and some limiting factors of discriminative training. We then present our simple Rival Training algorithm in section 3. Experimental results are given in section 4 for a continuous digit string and an isolated word recognition task. We summarize our main findings in section 5.

2. TRAINING CRITERIA

2.1. Maximum Likelihood Training

A commonly used training criterion to determine the parameters of the acoustic model is the Maximum Likelihood (ML) principle

\[ F_{ML}(\lambda) = \sum_{r=1}^{R} \log p_{\lambda}(X_r|W_r), \]

where \( X_r \) denotes a sequence of acoustic observation vectors for utterance \( r \in \{1, \ldots, R\} \), \( W_r \) denotes the corresponding sequence of spoken words and \( \lambda \) represents the set of acoustic model parameters. Maximum likelihood parameter estimation tries to maximize the likelihood for the spoken word sequence to generate the observed feature sequence. Since competing models are not taken into account, the recognition accuracy is maximized only indirectly.

2.2. MMI and Corrective Training

Several discriminative training (DT) approaches have been suggested [1, 2, 3] trying to maximize class separability and thus to improve recognition accuracy directly. A commonly used approach is to maximize

\[ F_{MMI}(\lambda) = \sum_{r=1}^{R} \log p_{\lambda}(W_r|X_r) \]

\[ = \sum_{r=1}^{R} \log \frac{p_{\lambda}(X_r|W_r) p(W_r)}{p_{\lambda}(X_r|W_{gen})} \]  \[ \tag{2} \]

with

\[ p_{\lambda}(X_r|W_{gen}) := \sum_W p_{\lambda}(X_r|W) p(W). \]  \[ \tag{3} \]

This training criterion ("Maximum Mutual Information", MMI) simultaneously tries to increase the likelihood of the
spoken word sequence $W$, and to decrease the likelihood of competing hypotheses $W$ contained in the “general model" $W_{gen}$. In general, $W_{gen}$ is obtained by a recognition pass on the training data. The acoustic parameters $\lambda$ are then reestimated in an iterative procedure, involving the determination of $W_{gen}$ according to the new estimate of $\lambda$ and the re-application of the reestimation equations in each iteration step. The reestimation equations are given e.g. in [8].

The implementation of this training process can be greatly simplified if the general model (3) is restricted to the recognized text. Since in this case correctly recognized sentences cancel out in equation (2), only misrecognized sentences contribute to the training criterion. The resulting algorithm is called corrective training (CT). It can be formulated on the basis of forward-backward (FB) probabilities calculated for the spoken and the recognized text (together with the corresponding feature vectors), and can be implemented completely within the Viterbi formalism.

Corrective training, however, has its limits when the training error is very low, since then only few material is used for reestimation of the acoustic parameters, with increasing risk of overfitting. In this case, using e.g. the “full” MMI approach seems to be more advisable, since then every sentence is used for reestimation. On the other hand, in MMI training a set of competing word sequences is involved such that the FB probabilities of each competing word sequence $W \in W_{gen}$ are weighted according to the posterior probability of $W$ [8]. Thus, for an MMI implementation, the simple Viterbi framework (assuming the FB probabilities to be either 1 or 0) is not sufficient.

This is also true for an implementation of MCE and “falsifying training” [9], since the observation vectors are weighted with an additional sigmoid smoothing function $f$.

3. RIVAL TRAINING

The goal of our work is a discriminative training algorithm that extends the set $M_r$ of competing word sequences used for discrimination (as compared to CT), but that can be implemented completely within the Viterbi framework. This is achieved by adding also correctly recognized sentences to $M_r$, using the second best hypothesis (the “rival”) as competing model ([6], p. 116). This amounts to defining $W_{gen}$ to be the best scored incorrect hypothesis. The goal of the training criterion (2) then is to increase the likelihood difference between the spoken utterance and its rival.

It is certainly most important to use those (correctly recognized) sentences with a low likelihood difference to the rival for reestimation, whereas sentences recognized with high confidence (i.e. with a large score difference to the rival) may even be skipped. In [6] and [9] this is achieved by introducing a sigmoid weighting function $f$. Since a continuous weighting function results in a weighting of observation vectors, preventing an implementation within the Viterbi formalism, we instead choose a “hard” threshold criterion to select the sentences used for discrimination:

- For correctly recognized sentences, use the second best hypothesis (the rival) for discrimination, if its score difference to the first best hypothesis is below a predefined score threshold.
- For misrecognized sentences, use the recognized utterance for discrimination (as in CT).

Apart from the redefinition of $W_{gen}$, the algorithm proceeds exactly in the lines of the implementation of corrective training presented in [8]. We call this approach “rival training” (RT).

We determine the score threshold by prefixing the fraction of correctly recognized utterances which — in addition to the misrecognized utterances — shall be used for discrimination. The score threshold can then be calculated as the corresponding quantil value of a histogram monitoring (for correctly recognized sentences) the score difference between the second best and the first best hypothesis. A quantil value of 0 corresponds to corrective training. A quantil of 0.2 means that 20% of the correct utterances are considered for discrimination (in addition to the misrecognized sentences). We calculate the score threshold once (in the first iteration) according to the given quantil value and then keep it constant during the subsequent iterations. This is motivated by the fact that there will be no reestimation of acoustic parameters anymore, if all training sentences are recognized correctly with the score difference to the rival exceeding the fixed score threshold value. We however obtained similar performance gains calculating the score threshold from the quantil value in each iteration of rival training.

4. RESULTS

Experiments were performed on two different tasks: continuous digit string recognition and isolated word recognition. In both cases, we use continuous Gaussian mixture emission distributions with a globally pooled variance vector.

The general training scheme proceeds as follows: First, we apply ML training using Viterbi alignment and the maximum approximation. In the continuous digit string task, the word penalty is optimized on the training corpus. References and word penalty are used for the baseline (ML) evaluation on the test corpus, and as a starting point for corrective and rival training. CT and RT are performed by iterating the reestimation equations given in [8]. The iteration factor $h$ controlling the step size of the iteration process was set to $h = 1.2$. In CT, in each iteration step we use the recognized text, obtained by a recognition pass on the training corpus, as the general model $W_{gen}$. In RT, in each iteration step we determine the “rivalizing text" from a 2—Best—List (obtained by a recognition pass) and the spoken text according to the algorithm presented in section 3. In the isolated word task the recognition pass of each iteration is constrained to a recognition on an input lattice, generated by an initial unconstrained recognition
pass using the ML references [10]. After a predefined number of iterations of CT or RT, respectively, we select the references yielding the lowest error rate on the training corpus for evaluation. In the digit string recognition task, the word penalty is again optimized on the training corpus.

4.1. Digit strings

In the first series of experiments, we used the male part of the SiETiLL corpus [4] for telephone line recorded German connected digit strings. This corpus consists of about 23k spoken digits in 7k sentences (190 speakers, about 2.5h of speech) for both training and test. The acoustic model consists of whole word HMMs (11 models including “zwo” as synonyme for the digit “2”) plus one silence model. Experiments are reported with 7 split references (30k densities, 118 per mixture). We used 11 cepstral features plus derivatives of the first 9 coefficients. The sampling rate was 8kHz, the frame shift 16ms. We applied cepstral mean subtraction on the sentence level and a LDA transform resulting in a 24 component feature vector.

Figure 1 shows the minimal word error rate (WER) on the training corpus obtained after 20 iterations of CT (quantil 0.0) and RT and the corresponding evaluation results on the test corpus. Although CT leads to a smaller WER on the training corpus, rival training gives a much better performance on the test data and reduces the WER from (1.85±0.18)% (ML) to 1.67% (10% relative improvement). CT, on the other hand, resulted in a WER of 1.89%, indicating overfitting\(^4\). It can be followed that by extending the set of sentences used for discrimination the generalization properties can be significantly improved. Figure 1 shows that the dependence of the recognition results on medium quantil values (0.25 ≤ quantil ≤ 0.6) is rather weak.

For larger quantil values, however, the test error rates slightly increase again. Experiments with 2 split references (which will be reported elsewhere) revealed that the performance on the test corpus decreases again for quantil values ≥ 0.5. We suppose that this is a matter of quality of the competing models rather than of quantity: It does not seem to be advisable to use utterances for discrimination which are already well discriminated, in terms of acoustic scores, from the spoken text. The quantil parameter leads to a selection of those utterances for discrimination which are most confusable with the spoken text in terms of acoustic scores.

4.2. Isolated words

For our experiments on an isolated word recognition task, we extracted a training and a test corpus, consisting of city and company names, first and family names and other single words, from SpeechDat\(^\text{®}3\) German material. The training corpus consists of 18k utterances (about 4h of speech), the test corpus of 10k utterances (about 3.5h of speech). The acoustic model is made of 16k triphones (34k densities, 61 densities per mixture). We use the same 10k lexicon for evaluation on the test corpus and for generation of the input lattice used to constrain the recognition passes during discriminative training. Signal analysis was done with a sampling rate of 8kHz and a frame shift of 10ms. We used 14 cepstral features plus derivatives, and applied recursive cepstral mean subtraction and a LDA transform resulting in a 36 component feature vector.

The training process is shown in figure 2; results on the test data are given in table 1.

Table 1 shows that rival training in all cases leads to significant better performance on the test data than corrective training. In particular, RT with quantil 0.5 reduces the ML baseline WER already after 10 iterations by 12% relative. Further iterating, the WER reveals fluctuations on the test corpus, which are, however, not statistically significant. Still, it has to be investigated in future research.
<table>
<thead>
<tr>
<th></th>
<th>qu.</th>
<th>it.</th>
<th>WER</th>
<th>it.</th>
<th>WER</th>
<th>it.</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td>0.0</td>
<td>10</td>
<td>7.80</td>
<td>15</td>
<td>7.68</td>
<td>20</td>
<td>7.80</td>
</tr>
<tr>
<td>RT</td>
<td>0.5</td>
<td>10</td>
<td>6.99</td>
<td>15</td>
<td>7.22</td>
<td>20</td>
<td>7.03</td>
</tr>
<tr>
<td>RT</td>
<td>1.0</td>
<td>10</td>
<td>7.54</td>
<td>15</td>
<td>6.95</td>
<td>20</td>
<td>7.20</td>
</tr>
</tbody>
</table>

Table 1: Results on the evaluation corpus (isolated word task) for maximum likelihood (ML), corrective (CT) and rival training (RT), for different numbers “it.” of training iterations. “qu.” denotes the quantil value.

if this is a result of overtraining.

Note that in nearly 40% of the utterances the spoken word was — even for a 10k lexicon and reasonable pruning parameters — discriminative enough so that the resulting N-Best list generated during recognition consisted of a single hypothesis. These utterances are skipped for the calculation of the score threshold and do not contribute to rival training (since no competing model can be generated). Thus, the interpretation of the quantil value refers to the remaining 60% of the utterances, i.e. a quantil value of 1.0 in this case means that all 11k utterances which produced more than a single hypothesis during recognition are used for discrimination, independent of the acoustic scores.

Figure 2 shows that the convergence of the training process is much faster for a smaller quantil value. As long as there is enough material for discrimination (to avoid overfitting to a small subset of training utterances) this may be used to accelerate the training process. A more detailed analysis reveals that even for quantil values smaller than 1.0 the set $M_r$ of utterances selected for discrimination is extremely fixed: In the experiment with quantil 0.5, more than 99% of the utterances that contribute to $M_r$ after 19 iterations also contribute to $M_r$ in the first iteration. This suggests that the whole training process could be restricted to the set of (about 5500) utterances that were used for discrimination in the first iteration of RT (corresponding to 30% of the 18k original training utterances, selected by the score threshold criterion triggered by the quantil value). This may further accelerate the training process.

5. CONCLUSIONS

We suggested a simple extension of the corrective training algorithm by using — in addition to misrecognized sentences — also a set of correctly recognized sentences for discrimination. In this case, the second best hypothesis (the “rival”) is used for discrimination, if its score (compared to the score of the spoken utterance) is below a predefined score threshold, triggered by a quantil parameter. We call this approach “Rival Training”. The motivation of this “hard” selection scheme is that it allows an easy implementation, based on the Viterbi formalism (in contrast to lattice-based discriminative methods like MMI and MCE which involve a weighting of observation vectors). In a continuous digit string and an isolated word recognition task we showed that rival training significantly increases recognition performance compared to maximum likelihood and corrective training even for mixture density models (whole words, triphones) with more than 60 densities per mixture.

Using more material for discrimination than compared to corrective training improves evaluation results, as long as the competing models are “close” (in terms of acoustic scores) to the spoken utterance. Moreover, if enough material is used for discrimination, the hard selection scheme can be used to accelerate the training process: first due to a faster convergence and second by a unique selection of training utterances (in the first iteration) to which the training process can be restricted. For practical applications it is important to note that the dependence of the performance on the quantil value was found to be rather weak: similar performance gains are observed using about 20% to 50% of the utterances for discrimination. This means that a fine-tuning of the quantil parameter is not necessary.

6. REFERENCES