Investigations on discriminative training in large scale acoustic model estimation

Janne Pylkkönen

Adaptive Informatics Research Centre, Helsinki University of Technology, Finland

janne.pylkkonen@tkk.fi

Abstract

In this paper two common discriminative training criteria, maximum mutual information (MMI) and minimum phone error (MPE), are investigated. Two main issues are addressed: sensitivity to different lattice segmentations and the contribution of the parameter estimation method. It is noted that MMI and MPE may benefit from different lattice segmentation strategies. The use of discriminative criterion values as the measure of model goodness is shown to be problematic as the recognition results do not correlate well with these measures. Moreover, the parameter estimation method clearly affects the recognition performance of the model irrespective of the value of the discriminative criterion. Also the dependence on the recognition task is demonstrated by example with two Finnish large vocabulary dictation tasks used in the experiments.

Index Terms: speech recognition, discriminative training, acoustic models, parameter estimation, lattice segmentation

1. Introduction

Discriminative training is today the de-facto standard for training high quality large scale speech recognition systems. Whereas maximum likelihood (ML) training aims at finding models that best match the observed distributions, discriminative methods define more explicit goals related to the recognition process itself. Discriminative methods differ in the way they address the goodness of the models. Maximum mutual information (MMI) criterion [1] still considers distributions, but uses conditional probabilities of the speech classes instead of probabilities of observations as in ML. A more recent minimum phone error (MPE) criterion [2] formulates a more direct expression for the expected phone error of the training sentences.

This paper investigates the sensitivity of these two common discriminative training criteria to two training aspects: lattice segmentation and parameter estimation method. For both aspects two strategies have been implemented and different combinations are applied to MMI and MPE training. The performance of the resulting models are tested on two Finnish large vocabulary dictation tasks.

The next two sections review different parameter estimation methods and lattice handling strategies. Then experimental results are presented, after which discussion and conclusions follow.

2. Parameter estimation

HMM parameter estimation can be seen as an optimization problem. Given a function that measures the goodness of the model, one aims to optimize it by changing the parameters of the HMMs. Because of the large scale of the optimization problem, few off-the-self algorithms are applicable to the optimization of acoustic models. The most widely used parameter estimation method for discriminative training, the extended Baum-Welch algorithm, is an extension of the traditional ML model estimation, specifically tuned for the given purpose.

There are not many comparisons between different parameter estimation methods of discriminative training in the literature. An early experiment tested a few methods in the context of MMI training of continuous phoneme recognition system [3]. It chose the extended Baum-Welch method as the best one, but the result was based solely on comparing the MMI criterion values, not the recognition performance. We will see in Sections 4 and 5 that when comparing parameter estimation methods, it is important to evaluate the actual recognition performance.

A more recent comparison found out it was possible to outperform the baseline extended Baum-Welch algorithm in discriminative HMM parameter estimation using the constrained line search algorithm [4].

2.1. Extended Baum-Welch

The most commonly used method for estimating HMM parameters under discriminative criterion is the extended Baum-Welch (EBW) algorithm [1]. The method uses re-estimation formulae that can be seen as extensions to the Baum-Welch algorithm used for ML parameter estimation. However, several heuristics are needed in order to get EBW to work well. Most notably, the method is rather sensitive to the smoothing parameter which determines the speed of convergence. Also directly applying EBW to mixture weight estimation is not feasible. Woodland and Povey [1] have presented now widely applied improvements to the original algorithm. This works also as the baseline estimation method for the experiments in this paper.

To promote generalization of discriminatively trained acoustic models, it is beneficial to smooth the EBW statistics with ML ones. This has been noted to be very important with the MPE criterion, for which the I-smoothing technique has been proposed to guarantee good generalization [2].

2.2. Gradient based methods

The gradient descent algorithm in its simplest form is a very straightforward optimization method that can be applied to any smooth and differentiable function. However, gradient methods may suffer from low convergence speed due to difficulty of setting a proper step size. Schlüter *et al.* [5] have demonstrated that in the context of discriminative HMM training the gradient method can be made to resemble EBW algorithm by using a step size that depends on certain model statistics. This way, results comparable to EBW can be obtained.

A more recent method derived from the gradient perspec-

tive is the constrained line search (CLS) algorithm [4]. Instead of using specific step sizes it uses some approximations to solve the critical points where derivatives are zero. The algorithm also explicitly assigns constraints that ensure the feasibility of the new estimates. In experiments the CLS has compared very well with the EBW.

2.3. Second-order methods

Second-order optimization methods use information about the Hessian matrix to find a better search direction than the plain gradient direction and can thus speed up the optimization. In larger optimization tasks the Hessian matrix becomes intractably large and some means to circumvent the computation and storing of the full matrix are needed.

One of the simplest second-order methods is the Quickprop algorithm, which uses a rough approximation of the diagonal of the Hessian. It has been used successfully at least in the context of minimum classification error criterion [6]. A more refined way of applying the second order information is used in conjugate gradient methods, which speed up the gradient search by taking into account the previous search direction and conjugacy of the gradients. Nevertheless, conjugate gradient methods have not been popular among speech recognition field.

One well-known set of second-order methods is the quasi-Newton family, including methods such as BFGS. These methods have some advantages over the conjugate gradient methods relating to the convergence speed and the accuracy of the line search [7]. Although the traditional implementation requires explicitly storing the Hessian matrix, there exists a rather simple limited-memory version that avoids this requirement [8].

For the current experiments the limited-memory BFGS (lBFGS) was chosen as the alternative HMM parameter optimization method due to its reported good performance. In preliminary testings it was noted that the maximum allowed parameter change at any optimization step needed to be limited. Otherwise some HMM parameters were prone to obtain unfeasible values and wreck the overall performance of the model. This is actually an issue that the constrained line search method [4] addresses directly. However, in this work a simpler constraint was applied, namely allowing maximum of 20% relative change for each parameter. The weights of one mixture were considered in a group as well as the means and covariances of one Gaussian when changing the optimization steps of these groups separately to meet the required limits.

2.4. Constraining the parameters

Usually the optimization algorithms are formulated in an unconstrained setting where parameters are free to obtain any real value. However, the HMM parameters must represent a proper probability distribution, and must therefore fulfill certain constraints. Specifically, the covariances must be positive definite and mixture weights of one mixture must sum to one.

In estimation methods developed for HMMs these constraints can be taken into account directly. In EBW the smoothing constant is determined so that covariances remain valid, and a special iterative estimation for mixture weights is used to ensure the sum-to-one constraint [1]. In CLS the explicit quadratic constraints avoid the possibility of improper covariances. For the mixture weights, the Lagrange multiplier method is used to constraint their estimation [4].

For a general purpose optimization method the easiest ways to apply constraints are penalty functions and parameter transformations. In the present work the latter was used. For optimization purposes the diagonal covariance parameters are transformed according to

$$\tilde{\sigma}_{i,k} = \sqrt{\sigma_{i,k} - MV}, \tag{1}$$

where $\sigma_{i,k}$ is the variance of the i:th dimension of Gaussian component k, and MV is the minimum allowed variance. An inverse transformation is applied after the optimization step to decode the covariance values. Mixture weights need to sum to one, so a soft max scaling is applied:

$$\tilde{w}_{i,m} = \log w_{i,m}, \tag{2}$$

$$w_{i,m} = \frac{\exp \tilde{w}_{i,m}}{\sum_j \exp \tilde{w}_{j,m}}. \tag{3}$$

These transformations need to be taken into account in the gradient computations as well.

3. Lattice segmentation

In order to address the recognition problems, discriminative training methods need to do similar kind of recognition of the training material as is done in the target system. The full recognition is usually too time consuming so an approximate recognition model is used instead. Lattices have been found to give good results with discriminative training [1] as they effectively restrict the available hypotheses during the training but are loose enough to allow a realistic model of the recognition.

Although lattices already restrict the search space for state segmentation, it is usually necessary to further limit the search to reduce the computational load. Beam pruning is an easy and effective way to ignore the improbable hypotheses. It is also well justified as similar technique is used also in the actual recognition. Even more strict pruning is possible by storing time stamps of an initial segmentation to the lattice and using these to guide the search. In the extreme case, the state occupancies are kept fixed during the training.

It has been previously noted that resegmenting the lattices during training slightly improves the results [1]. Therefore, for the current paper the least restrictive lattice segmentation strategy was sought. A full Baum-Welch (BW) segmentation was implemented that applies only beam pruning and no other approximations. The segmentation is recomputed at every iteration.

Unfortunately, the straightforward implementation of the Baum-Welch segmentation poses some problems in the discriminative training. To simulate the effect of the language model scaling used in recognition and to increase the number of competing hypotheses, discriminative training algorithms use acoustic scaling to compress the likelihood values of the acoustic models [1]. The easiest way to accomplish this is to scale the acoustic probabilities framewise and use them otherwise as normally. This, however, has the adverse result that the likelihoods of different state segmentations of the same hypothesis no longer sum up properly if they are to be considered as a single hypothesis. This is the case, for example, in the numerator of the MMI criterion.

To circumvent the problem, a second lattice handling strategy was experimented in this paper. The algorithm, referred to as Multipath Viterbi (MPV), considers only the best state segmentation of each hypothesis and therefore avoids the problem with likelihood summing. Using lattices preprocessed with FST-tools [9] this kind of algorithm becomes fairly simple.

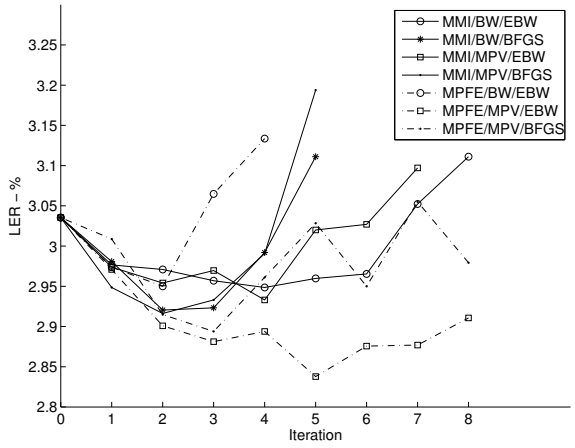

Figure 1: *Results of Speecon task over the development set.*

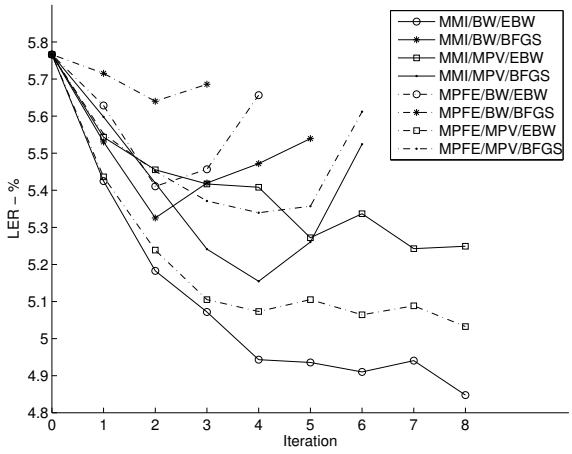

Figure 2: *Results of SpeechDat task over the development set.*

Something similar can be accomplished using Viterbi segmentation over fixed time alignments [1]. In some cases this even improved results over less restrictive segmentations. Another benefit from using Viterbi segmentations of hypotheses is that this prevents the smearing of phoneme transitions that can otherwise be problematic when acoustic scaling is used.

4. Experiments

In the experiments different combinations of lattice segmentation and parameter estimation strategies were tested with MMI and MPE criteria in two Finnish large vocabulary dictation tasks.

The simpler of the tasks was based on the Finnish Speecon database [10], from which 15 hours of clean speech (sample rate 16kHz) from 310 speakers were used for training. Separate development and test sets were used with 40 distinct speakers and about 1.9 hours of speech in both sets.

A slightly harder task was from the Finnish SpeechDat database[1], which consists of 4000 speakers recorded over fixed telephone line. For that corpus, 55 hours from 3696 speakers were used for training, and for both development and test sets 150 separate speakers were allocated, both containing about 2.2 hours of speech. Both tasks were dictation tasks, only read sentences were used for training and testing. The latter task was harder not only because of the lower sound quality, but also due to the wider range of speakers, including children. No adaptation was performed.

The experiments were run using a Finnish morph-based large vocabulary speech recognition system [11]. In both tasks the acoustic features were 39 dimensional MFCCs with cepstral mean subtraction. The Gaussians had diagonal covariances and a global diagonalising transform was used. In the Speecon task the models contained 26126 Gaussians in 1186 mixtures. This corresponds to the average of 255 frames in the training set for each Gaussian. In the SpeechDat task the models had 85758 Gaussians in 1783 mixtures, equating 290 frames per a Gaussian on average.

For MPE a frame based implementation was used, which has been shown to give good results [12, 13]. The implementation details followed mostly the MPFE-nosil method described by Povey [13], and the suggested I-smoothing value of 400 was used. The numerator was segmented with Viterbi segmentation also in BW case to have a unique reference for the phone accuracy function. However, the preliminary experiments showed that when using the MMI criterion with BW lattice segmentation it was crucial to have also the numerator lattice use full BW instead of Viterbi.

[1] http://www.speechdat.org/

Table 1: *Final test set results of Speecon task in LER. The baseline ML model had a LER of 3.33%, which corresponds to word error rate of 13.7%.*

method	EBW	lBFGS
MMI/BW	3.15%	3.26%
MMI/MPV	3.12%	3.24%
MPFE/BW	3.30%	-
MPFE/MPV	3.02%	3.02%

For simplicity, I-smoothing was not used for any other case than MPFE with EBW estimation. Specifically, the MPFE was tested with lBFGS optimization without I-smoothing to see the performance of the discriminative criterion in its simplest form. Besides the maximum parameter change limitation discussed in Section 2.3, the lBFGS method only needs an initial step size, which was tuned in the preliminary tests to perform a step comparative to that of further iterations.

The parameter estimation methods were run at most 8 iterations in each case, and were in some cases stopped earlier if clear overtraining was observed. The results of the experiments were measured with letter error rate (LER) to have a finer resolution for the results. This measure is well suitable for Finnish in which words are relatively long and consist of several morphs. The language model used morphs as its units [11].

Figure 1 shows the results of the Speecon task over the development set at each iteration of the model estimation. The MPFE/BW combination was not tested with the lBFGS estimation due to low performance of the combination in other settings. The figure shows clearly that in this task all but the MPFE/MPV combination with EBW parameter estimation exhibit overtraining. The best performing iteration for each method was selected according to these results and the test set was then recognized using these models. The test set results are shown in Table 1.

Figure 2 shows the development set results for the Speech-

Table 2: *Final test set results of SpeechDat task in LER. The baseline ML model had a LER of 7.16%, which corresponds to word error rate of 22.3%.*

method	EBW	lBFGS
MMI/BW	6.05%	6.56%
MMI/MPV	6.47%	6.43%
MPFE/BW	6.78%	-
MPFE/MPV	5.95%	6.35%

Dat task. Again, the lBFGS methods exhibit overtraining, but now using EBW estimation leads to nice convergence except for the MPFE/BW combination. The test set results are in Table 2. Note that although MMI/BW/EBW model was the best in the development set, MPFE/MPV/EBW is the best using the separate test set.

5. Discussion

In the simpler Speecon task the differences between the results of the different training strategies were not that large. Most notably, the lBFGS estimation performed worse in MMI case than the EBW and the BW segmentation failed to work with the MPFE criterion. The best models were obtained with the MPFE/MPV combination, for which both EBW and lBFGS obtained equally good models.

The more difficult recognition task, SpeechDat, shows more variation between the performances of different training strategies. In this task EBW was better than lBFGS except with the MMI/MPV combination, which was still worse than the MMI/BW combination using the EBW estimation. It is interesting that the best MMI model used Baum-Welch lattice segmentation, whereas for MPFE the Multipath Viterbi segmentation was clearly the best. The suitability of MPV style lattice segmentation for MPFE training is emphasized by the similar results in the Speecon task. This suggests that with a discriminative criterion such as MPE that closely matches the recognition process, also the lattice segmentation should be done in Viterbi style as is done in recognition.

Another important observation is that the values of the discriminative criteria do not correlate with the recognition performance of the models. In all cases, lBFGS was able to optimize the criteria better than EBW, but only in the Speecon task with MPFE criterion it was able to obtain equal recognition performance. Surprisingly, in the SpeechDat task the EBW estimation even decreased the MMI measure with the MMI/BW combination, although the resulting model was one of the best ones. This anomaly may be due to problems in handling acoustic scaling in Baum-Welch style lattice segmentation.

In the current implementations overtraining was much more severe with lBFGS than with EBW. However, this shouldn't confuse the results as a development set was used to pick the best model in each case. For lBFGS, a decreasing limit in the maximum allowed parameter change might be used to alleviate this problem as was done in Liu *et al.* [4].

The model complexity in the experiments was rather high, but this was because the complexity was optimized for the best baseline ML model performance. Generally the discriminative methods work better if more training data is available, so with that respect the rather large improvements (9.3% and 16.9% relative improvements over the ML model in Speecon and SpeechDat tasks, respectively) show very good performance of the best tested discriminative training methods.

6. Conclusions

Lattice segmentation may have a large impact on the performance of discriminatively trained acoustic models. A full Baum-Welch segmentation seems to work well with MMI training, whereas a Viterbi style segmentation of different hypotheses was better for MPE training.

The commonly used MMI and MPE criteria are not good in a way that simply optimizing them is not guaranteed to lead to good results. The extended Baum-Welch algorithm used to estimate discriminative acoustic models interacts favorably with the training criteria and is able to obtain better performing models than a general purpose optimization algorithm.

7. Acknowledgements

The author wishes to thank the Graduate School of Language Technology in Finland, the Emil Aaltonen foundation and KAUTE foundation for funding the research.

8. References

[1] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech and Language*, vol. 16, pp. 25–47, 2002.

[2] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002, pp. 105–108.

[3] S. Kapadia, V. Valtchev, and S. J. Young, "MMI training for continous phoneme recognition on the TIMIT database," in *Proc. ICASSP*, 1993, pp. 491–494.

[4] P. Liu, C. Liu, H. Jiang, F. Soong, and R.-H. Wang, "A constrained line search optimization method for discriminative training of hmms," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 900–909, 2008.

[5] R. Schlüter, W. Macherey, B. Müller, and H. Ney, "Comparison of discriminative training criteria and optimization methods for speech recognition," *Speech Communication*, vol. 34, pp. 287–310, 2001.

[6] E. McDermott, T. J. Hazen, J. L. Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large-vocabulary speech recognition using minimum classification error," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 203–223, 2007.

[7] D. F. Shanno, "Conjugate gradient methods with inexact searches," *Mathematics of Operations Research*, vol. 3, no. 3, pp. 244–256, 1978.

[8] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, pp. 503–528, 1989.

[9] L. Hetherington, "The MIT finite-state transducer toolkit for speech and language processing," in *Proc. Interspeech*, 2004, pp. 2609–2612.

[10] D. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling, "SPEECON - speech databases for consumer devices: Database specification and validation," in *Proc. LREC*, 2002, pp. 329–333.

[11] T. Hirsimäki, J. Pylkkönen, and M. Kurimo, "Importance of high-order N-gram models in morph-based speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 724–732, 2009.

[12] J. Zheng and A. Stolcke, "Improved discriminative training using phone lattices," in *Proc. Interspeech*, 2005, pp. 2125–2128.

[13] D. Povey and B. Kingsbury, "Evaluation of proposed modifications to MPE for large scale discriminative training," in *Proc. ICASSP*, 2007, pp. 321–324.