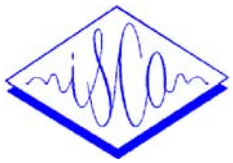

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

COMPARISON OF OPTIMIZATION METHODS FOR DISCRIMINATIVE TRAINING CRITERIA

R. Schlüter, W. Macherey, S. Kanthak, H. Ney, L. Welling
Lehrstuhl für Informatik VI
RWTH Aachen – University of Technology
D-52056 Aachen, Germany
Email: schlueter@informatik.rwth-aachen.de

ABSTRACT

In this work we compare two parameter optimization techniques for discriminative training using the MMI criterion: the extended Baum-Welch (EBW) algorithm and the generalized probabilistic descent (GPD) method. Using Gaussian emission densities we found special expressions for the step sizes in GPD, leading to reestimation formula very similar to those derived for the EBW algorithm. Results were produced for both the *TI digitstring* and the *SieTill* corpus for continuously spoken American English and German digitstrings. The results for both techniques do not show significant differences. This experimental results support the strong link between EBW and GPD as expected from the analytic comparison.

1. INTRODUCTION

In an increasing number of applications discriminative training criteria such as *Maximum Mutual Information* (MMI) [8, 12, 15] and *Minimum Classification Error* (MCE) [2, 5, 15] have been used. In MCE training, an approximation for the error rate on the training data is optimized, whereas MMI optimizes the *a posteriori* probability of the training utterances and hence the class separability.

It has been shown that discriminative training criteria are able to produce significant improvements in word error rate in comparison to the conventional *Maximum Likelihood* (ML) training criterion. Since there does not exist any discriminative training method which is guaranteed to converge under all practical conditions, much effort has been made to develop parameter optimization techniques with fast and reliable convergence.

Here two parameter optimization techniques for discriminative training, the *Extended Baum Welch* (EBW) algorithm [12] and the *Generalized Probabilistic Descent* (GPD) [2, 5, 15], will be discussed. EBW is an extension to the standard *Baum Welch* algorithm designed for optimization of the MMI criterion. GPD, which is commonly used for MCE, essentially performs a gradient descent on the discriminative training criterion and hence is easily transferred to other criteria like MMI.

In this work the MMI criterion is applied to train connected digit recognizers. The parameter optimization methods EBW and a special form of GPD will be compared analytically. Experimental results are presented on the *TI digitstring* and the *SieTill* corpus applying corrective training [12].

2. DISCRIMINATIVE TRAINING

The training data shall be given by $r = 1, ..., R$ training utterances, each consisting of a sequence X_r of acoustic observation vectors $x_{r,1}, x_{r,2}, ..., x_{r,T_r}$ and the corresponding sequence W_r of spoken words $w_{r,1}, w_{r,2}, ...w_{r,N_r}$. The *a posteriori* probability for the word sequence W_r given the acoustic observation vectors X_r shall be denoted by $p_\lambda(W_r|X_r)$. Similarly, $p_\lambda(X_r|W_r)$ and $p(W_r)$ represent the emission and language model probabilities for the acoustic observation sequence X_r and the word sequence W_r. In the following, the language model probabilities are supposed to be given. Hence the parameter λ represents the set of all parameters of the emission probabilities $p_\lambda(X_r|W_r)$.

Then the MMI criterion, which is defined by the sum over the logarithms of the *a posteriori* probabilities of each training utterance, is given by:

$$\begin{aligned} F(\lambda) &= \sum_{r=1}^{R} \log p_\lambda(W_r|X_r) \\ &= \sum_{r=1}^{R} \log \frac{p(W_r)p_\lambda(X_r|W_r)}{\sum_W p(W)p_\lambda(X_r|W)}. \end{aligned}$$

Clearly an optimization of the MMI criterion tries to simultaneously maximize the emission probabilities of the spoken training sentences and to minimize a weighted sum over the emission probabilities of each competing sentence given the acoustic observation sequence for each training utterance. The weights in the sum over the competing sentences are given by the language model probabilities relative to the spoken sentence. Thus the MMI criterion optimizes the class separability according to the words under consideration of the language model.

2.1. Parameter Optimization

Here only the case of single Gaussian densities with density specific variances will be discussed. Similar calculations hold for the more general cases of mixture densities with pooled, mixture or density specific variances.

2.1.1. Gradient Descent

One possibility to maximize the MMI criterion consists of a gradient descent with the following iterative reestimation formula for the parameters:

$$\overline{\lambda} = \lambda + \epsilon \cdot \frac{\partial F(\lambda)}{\partial \lambda}. \qquad (1)$$

Now let $p(x|\lambda_s)$ be the emission probability of the acoustic observation vector x given an HMM state s, with λ_s

the parameters of the acoustic model in state s. Then the derivative of the MMI criterion with respect to parameters λ_s is given by:

$$\frac{\partial F(\lambda)}{\partial \lambda_s} = \, , _s \left(\frac{\partial \log p(x|\lambda_s)}{\partial \lambda_s} \right), \qquad (2)$$

where the discriminative averages $, _s$ are defined by:

$$, _s (g(x)) = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \left(\gamma_{r,t}(s;W_r) - \gamma_{r,t}^{gen}(s) \right) \; g(x_{r,t}). \quad (3)$$

These make use of the *Forward-Backward* (FB) probabilities of the spoken word sequence W_r:

$$\gamma_{r,t}(s;W_r) \; = \; p_\lambda(s_t = s|X_r, W_r), \qquad (4)$$

and the generalized FB probabilities for the sums over all competing word sequences W:

$$\begin{aligned} \gamma_{r,t}^{gen}(s) \quad &= \quad \sum_W p_\lambda(W|X_r) \; \gamma_{r,t}(s;W) \\ &= \quad p_\lambda(s_t = s|X_r). \end{aligned}$$

The generalized FB probability is simply a sum over the conventional FB probabilities of each competing sentence weighted by its posterior probability.

2.1.2. Extended Baum-Welch Algorithm

Discriminative training with the MMI criterion usually applies an extended version of *Baum Welch* training, the EBW algorithm [11, 12, 13]. There the MMI criterion is maximized via the following auxiliary function:

$$\begin{aligned} & S(\lambda, \overline{\lambda}) \\ &= \quad \sum_s \sum_{r=1}^{R} \sum_{t=1}^{T_r} \left[\gamma_{r,t}(s;W_r) - \gamma_{r,t}^{gen}(s) \right] \log p(x_{r,t}|\overline{\lambda}_s) \\ &\quad + \sum_s D_s \int dx \, p(x|\lambda_s) \log p(x|\overline{\lambda}_s), \end{aligned}$$

which is to be optimized iteratively. Differentiation with respect to the iterated parameters $\overline{\lambda}_s$ leads to the following expression, from which reestimation formulae can be derived:

$$\begin{aligned} \frac{\partial S(\lambda, \overline{\lambda})}{\partial \overline{\lambda}_s} \quad &= \quad , _s \left(\frac{\partial \log p(x|\overline{\lambda}_s)}{\partial \overline{\lambda}_s} \right) \\ &\quad + D_s \int dx \, p(x|\lambda_s) \frac{\partial \log p(x|\overline{\lambda}_s)}{\partial \overline{\lambda}_s}. \end{aligned}$$

2.1.3. Reestimation Formulae

Let the emission probabilities be given by single Gaussians with diagonal covariances. Then the reestimation formulae for initial mean and variance vectors for state s, μ_s and σ_s^2 are given as follows:

- **GPD**:

$$\begin{aligned} \hat{\mu}_{s,(\mathrm{GPD})} \quad &= \quad \mu_s + \frac{\epsilon_{\mu_s}}{\sigma_s^2} [, _s(x) - , _s(1) \mu_s] \\ \hat{\sigma}^2_{s,(\mathrm{GPD})} \quad &= \quad \sigma_s^2 + \frac{\epsilon_{\sigma_s}}{2\sigma_s^4} [, _s(x^2) - 2, _s(x) \mu_s \\ &\qquad + , _s(1) \cdot (\mu_s^2 - \sigma_s^2)]. \end{aligned}$$

- **EBW**:

$$\begin{aligned} \hat{\mu}_{s,(\mathrm{EBW})} \quad &= \quad \frac{, _s(x) + D_s \mu_s}{, _s(1) + D_s} \\ \hat{\sigma}^2_{s,(\mathrm{EBW})} \quad &= \quad \frac{, _s(x^2) + D_s(\sigma_s^2 + \mu_s^2)}{, _s(1) + D_s} - \hat{\mu}_s^2 \end{aligned}$$

Although there do exist proofs of convergence for both GPD [4] and EBW [3, 7], the step sizes needed to guarantee convergence are impractical by leading to very slow convergence [12]. In practice, faster convergence is achieved in the EBW case, if the iteration constants D_s are chosen such that the denominators in the reestimation equations and the according variances are kept positive:

$$D_s = h \cdot \max \left\{ D_{s,\min}, \frac{1}{\beta} - , _s(1) \right\}. \qquad (5)$$

Here $D_{s,\min}$ denotes an estimation for the minimal iteration constant which guarantees the positivity of the variance in state s, and the iteration factor $h > 1$ controls the convergence of the iteration process, high values leading to low step sizes. The constant $\beta > 0$ is chosen to prevent overflow caused by low-valued denominators.

2.1.4. Comparison GPD vs. EBW

A direct comparison of the reestimation formulae for GPD and EBW leads to the following special expressions for the iteration step sizes for GPD:

$$\epsilon_{\sigma_s} = 2\sigma_s^2 \epsilon_{\mu_s} = 2\sigma_s^4 \min \left\{ \frac{1}{, _s(1) + h D_{\min}}, \frac{\beta}{h} \right\}. \qquad (6)$$

Using Eq. 6 we find the reestimation formulae for GPD and EBW to be very similar:

$$\begin{aligned} \hat{\mu}_{s,(\mathrm{GPD})} \quad &= \quad \hat{\mu}_{s,(\mathrm{BW})} \\ \hat{\sigma}^2_{s,(\mathrm{GPD})} \quad &= \quad \hat{\sigma}^2_{s,(\mathrm{BW})} + (\mu_s - \hat{\mu}_{s,(\mathrm{BW})})^2. \end{aligned}$$

In addition this comparison shows that the choice of the iteration constant in the EBW case implies an upper bound of the resulting step size, which is given by the constant β/h.

2.2. Approximations

In the following experiments we use an approximation for the calculation of the generalized FB probabilities. The sum over all competing sentence hypotheses is typically evaluated using N-best lists or, especially for large vocabulary, word graphs produced by a preceding recognition pass over the training data. Here the competing model is reduced to the best recognized sentence only, such that the generalized FB probability could be replaced by the conventional FB probability for the best recognized sentence. As a consequence only misrecognized training sentences make a contribution to the optimization process. This method is called corrective training [12].

In addition, time alignment for calculation of the FB probabilities is performed using the Viterbi approximation [10].

3. RESULTS

Experiments were done for the recognition of continuous digitstrings using both the *TI digitstring* [9] corpus for American English digits and the *SieTill* [6] corpus for

telephone line recorded German digits. In Table 1 some information on corpus statistics is summarized.

Table 1. Corpus statistics for the *TI digitstring* and the *SieTill* corpus.

corpus		female		male	
		sent.	digits	sent.	digits
TI	test	4389	14424	4311	14159
	train	4388	14414	4235	13915
SieTill	test	6176	20205	6938	22881
	train	6113	20115	6835	22463

The recognition systems for both corpora are based on whole word HMMs using continuous emission densities. They are characterized as follows:

TI digitstring recognition system:

- single Gaussian densities using state dependent variance vectors
- gender dependent whole word HMMs for 11 English digits including 'oh' and gender dependent silence models
- per gender 357 states plus one state for silence
- 16 cepstral features with first and second derivatives.

SieTill recognition system:

- single Gaussian densities using a pooled variance vector
- gender dependent whole word HMMs for 11 German digits including '*zwo*' and gender dependent silence models
- per gender 223 states plus one state for silence.
- 12 cepstral features with first derivatives and the second derivative of the energy.

Both baseline recognizers apply ML training using the Viterbi approximation [10] and their results serve as starting points for the additional discriminative training. A detailed description of the baseline system could be found in [16].

Since discriminative training methods could not guarantee convergence under realistic conditions, we first investigated the convergence behaviour of the MMI criterion. Using iteration factors $h = 5$ for mixture specific (*TI digitstring*, cf. Fig. 1) and $h = 2$ for pooled variances (*SieTill*) we found relatively smooth convergence for both GPD and EBW.

Similar results could be observed for the word error rates on test and training data, as is shown in Fig. 2 for the male portion of the *TI digitstring* corpus. Clearly, convergence on test and training data is comparable and thus the convergence of the error rate on the training data could be used as criterion to stop an iteration.

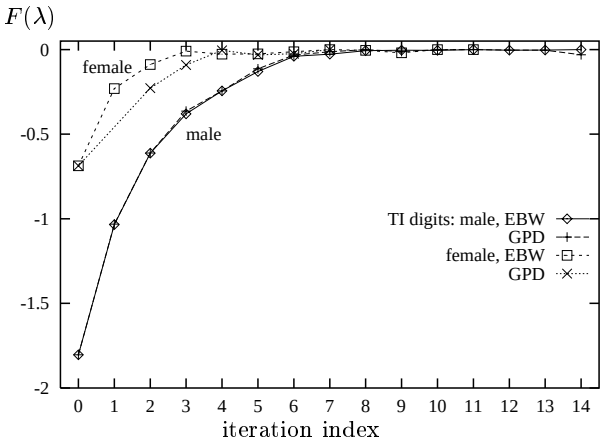

Figure 1. MMI criterion F for male speakers in the course of the iteration process.

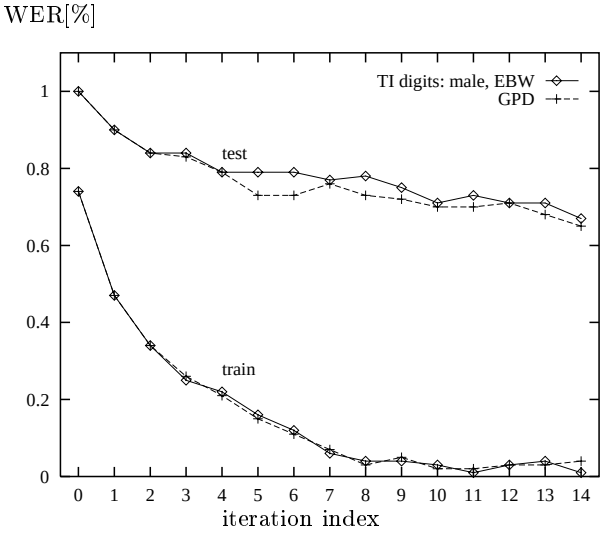

Figure 2. Evolution of the word error rate for male speakers in the course of the iteration process for the *TI digitstring* test and training corpus.

In the case of the *TI digitstring* corpus, an interesting fact is the reduction to no errors on the training data. On the one hand this shows the strong homogeneity of the *TI digitstring* corpus and that single densities should at least have the ability to model such a corpus completely without significant numbers of errors. On the other hand it clearly brings up the limitation of corrective training, since having no or very few errors on the training data prevents any further change in the iteration process.

As expected analytically, the recognition results on both *TI digitstring* and *SieTill* corpus do not show any significant difference for GPD and EBW reestimation. As could be seen in Tables 2 and 3, no consistent differences between results using GPD and EBW reestimation could be observed.

For the *TI digitstring* corpus the corrective MMI training removed all word errors on the training set. On the test data a relative improvement of more than 30% in word and sentence error rate was achieved.

On the *SieTill* corpus, the MMI training more than halfed the error rates on the training set and led to a relative improvement of 40% for the word and sentence error rate on the test set.

Table 2. Recognition results for the *TI digitstring* corpus.

corp.	method	del/ins/sub	WER[%]	SER[%]
train	ML	79/11/68	0.56	1.69
	EBW	0/0/0	0.0	0.0
	GPD	2/2/2	0.02	0.06
test	ML	56/31/120	0.72	2.00
	EBW	35/24/83	0.50	1.38
	GPD	36/24/75	0.47	1.32

Table 3. Recognition results for the *SieTill* corpus.

corp.	method	del/ins/sub	WER[%]	SER[%]
train	ML	449/189/1983	6.2	16.9
	EBW	249/185/683	2.6	7.5
	GPD	231/183/656	2.5	7.2
test	ML	621/324/2297	7.5	19.7
	EBW	445/318/1173	4.5	11.7
	GPD	419/322/1132	4.4	11.3

It should be noted that, using ML training, our recognition systems perform better with single Laplacians than with single Gaussians. The ML result for the sentence error rate on the *TI digitstring* corpus was 1.69%. Similarly, for the *SieTill* corpus the word error rate with ML trained single Laplacians was 6.1%. In comparison to these results, the error rates for MMI training with single Gaussians still outperform the ML training with single Laplacians at least by 20% relatively.

4. CONCLUSION

Two approaches for the optimization of discriminative criteria, the *generalized probabilistic descent* (GPD) and the *extended Baum-Welch* (EBW) algorithm were investigated. For the case of Gaussian densities, step sizes for the GPD algorithm were presented, showing strong similarities between GPD and EBW. Comparative experiments on both the *TI digitstring* and the *SieTill* corpus were done. In confirmation with the analytic results, the experimental results do not inidicate significant differences between GPD and EBW. Using single densities relative improvements of more than 30% on the *TI digitstring* and of 40% on the *SieTill* corpus in comparison to the initial ML results could be achieved.

Acknowledgement. This work was partly supported by Siemens AG, Munich.

REFERENCES

[1] C. M. Ayer, M. J. Hunt, D. M. Brookes. "A Discriminatively Derived Transform for Improved Speech Recognition," Proc. *1993 Europ. Conf. on Speech Communication and Technology*, Vol. 1, pp. 583-586, Berlin, September 1993.

[2] J. Bauer. "Enhanced Control and Estimation of Parameters for a Telephone Based Isolated Digit Recognizer," Proc. *1992 Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 2, page 1531-1534, Munich, April 1997.

[3] L. E. Baum, J. A. Eagon. "An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model for Ecology," Bulletin of the American Mathematical Society, Vol. 73, pp. 360-363, 1967.

[4] W. Chou, B.-H. Juang, C.-H. Lee. "Segmental GPD Training of HMM Based Speech Recognizer," Proc. *1992 Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, page 473-476, San Francisco, CA, March 1992.

[5] W. Chou, C.-H. Lee, B.-H. Juang. "Minimum Error Rate Training based on N-Best String Models," Proc. *1993 Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 652-655, Minneapolis, MN, April 1993.

[6] T. Eisele, R. Haeb-Umbach, D. Langmann, "A comparative study of linear feature transformation techniques for automatic speech recognition," in *Proc. Int. Conf. on Spoken Language Processing*, Vol. I, pp. 252-255, Philadelphia, PA, October 1996.

[7] P. S. Gopalakrishnan, D. Kanevsky, A. Nádas, D. Nahamoo. "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems," IEEE Transactions on Information Theory, Vol. 37, Nr. 1, pp. 107-113, January 1991.

[8] S. Kapadia, V. Valtchev, S. J. Young. "MMI Training for Continuous Phoneme Recognition on the TIMIT Database," Proc. *1993 Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 494-494, Minneapolis, MN, April 1993.

[9] R. G. Leonard. "A database for speaker-independent digit recognition," *Int. Conf. on Acoustics, Speech and Signal Processing 1984*, pp. 42.11.1-42.11.4, San Diego, CA, March 1984.

[10] H. Ney. "Acoustic Modeling of Phoneme Units for Continuous Speech Recognition," Proc. *Fifth Europ. Signal Processing Conf.*, pp 65-72, Barcelona, September 1990.

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

[12] Y. Normandin. "Maximum Mutual Information Estimation of Hidden Markov Models," *Automatic Speech and Speaker Recognition*, C.-H. Lee, F. K. Soong, K. K. Paliwal (eds.), pp. 57-81, Kluwer Academic Publishers, Norwell, MA, 1996.

[13] Y. Normandin, R. Lacouture, R. Cardin. "MMIE Training for Large Vocabulary Continuous Speech Recognition," Proc. *1994 Int. Conf. on Spoken Language Processing*, Vol. 3, pp. 1367-1370, Yokohama, September 1994.

[14] K. K. Paliwal, M. Bacchiani, Y. Sagisaka. "Minimum Classification Error Training Algorithm for Feature Extractor and Pattern Classifier in Speech Recognition," Proc. *1995 Europ. Conf. on Speech Communication and Technology*, Vol. 1, pp. 541-544, Madrid, September 1995.

[15] W. Reichl, G. Ruske. "Discriminative Training for Continuous Speech Recognition," Proc. *1995 Europ. Conf. on Speech Communication and Technology*, Vol. 1, pp. 537-540, Madrid, September 1995.

[16] L. Welling, H. Ney, A. Eiden, C. Forbrig. "Connected Digit Recognition using Statistical Template Matching," Proc. *1995 Europ. Conf. on Speech Communication and Technology*, Vol. 2, pp. 1483-1486, Madrid, September 1995.

[17] F. Wolfertstetter. *Verallgemeinerte stochastische Modellierung für die automatische Spracherkennung,* Ph.D. thesis, Lehrstuhl für Mensch-Maschine-Kommunikation, Technical University Munich, 1996.