COMPARISON OF DISCRIMINATIVE TRAINING CRITERIA

Ralf Schlüter and Wolfgang Macherey

Lehrstuhl für Informatik VI
RWTH Aachen – University of Technology
D-52056 Aachen, Germany
Email: schlueter@informatik.rwth-aachen.de

ABSTRACT

In this paper, a formally unifying approach for a class of discriminative training criteria including *Maximum Mutual Information* (MMI) and *Minimum Classification Error* (MCE) criterion is presented, including the optimization methods gradient descent (GD) and *extended Baum-Welch* (EB) algorithm. Comparisons are discussed for the MMI and the MCE criterion, including the determination of the sets of word sequence hypotheses for discrimination using word graphs. Experiments have been carried out on the *SieTill* corpus for telephone line recorded German continuous digit strings. Using several approaches for acoustic modeling, the word error rates obtained by MMI training using single densities always were better than those for Maximum Likelihood (ML) using mixture densities. Finally, results obtained for corrective training (CT), i.e. using only the best recognized word sequence in addition to the spoken word sequence, could not be improved by using the word graph based discriminative training.

1. INTRODUCTION

In an increasing number of applications discriminative training criteria such as *Maximum Mutual Information* (MMI) [6] and *Minimum Classification Error* (MCE) [1] 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. Based on [6], we present a formally unifying approach for a class of discriminative criteria including the MMI and the MCE criterion, thus extending a comparison done in [7]. In a previous study [9], we also found a unifying approach for the optimization methods gradient descent and *extended Baum-Welch* (EB) algorithm which was transfered to the unified criterion presented here.

Experimental results are presented for the *SieTill* corpus for telephone line recorded German connected digit strings. In order to investigate the abilities of discriminative training to improve ML training results, we performed comparative experiments for several approaches of acoustic modeling, such as single vs. mixture densities, pooled vs. state specific variances and an optional linear discriminant analysis (LDA).

Following previous studies [9], we also performed experiments comparing GD with EB optimization for MMI training of mixture densities showing no significant differences. Furthermore, for determining the sets of competing word hypotheses for discrimination, we performed experiments using CT [6], or word graphs for efficient representation of all competing word hypotheses. These experiments were initialized with our best results using CT, where only the best recognized word sequence is used for discrimination. We did not observe further improvements in word error rate, although in case of the use of word graphs a further convergence of the criterion was found.

2. DISCRIMINATIVE TRAINING

The training data shall be given by training utterances $r = 1...R$, each consisting of a sequence X_r of acoustic observation vectors $x_{r1}, x_{r2}, ..., x_{rT_r}$ and the corresponding sequence W_r of spoken words. 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 according emission and language model probabilities respectively. 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 $p_\lambda(X_r|W_r)$. Finally, let $\mathcal{M}_r$ denote the set of word sequences, which are considered for discrimination in utterance r. A class of discriminative training criteria F_D including MMI and MCE could then be defined by the expression

$$F_D(\lambda) = \sum_{r=1}^{R} f\Big(\log \frac{p^\alpha(W_r)p_\lambda^\alpha(X_r|W_r)}{\sum_{W \in \mathcal{M}_r} p^\alpha(W)p_\lambda^\alpha(X_r|W)}\Big).$$

The choice of the exponent α, the smoothing function f and the set $\mathcal{M}_r$ of word sequences for discrimination decide which criterion is represented. In particular, choosing $\alpha = 1$ and $f(\xi) = \xi$ yields the MMI criterion. On the other hand, using the sigmoid function $f(\xi) = -1/\left[1 + exp(2\beta\xi)\right]$ yields an equivalent version of the MCE criterion, which is to be maximized. Ideally, in case of the MMI criterion the set $\mathcal{M}_r$ would contain all possible word sequences. In practice, $\mathcal{M}_r$ is obtained through a recognition pass and is represented by N-best lists or word graphs. For MCE the spoken word sequence has to be excluded from this set. The contribution of each competing sentence to reestimation is controlled by the exponent α, where very large values of α lead to a maximum approximation. For the MMI criterion the latter is called corrective training (CT), where only the best recognized word sequences are used for discrimination. The smoothing function f leads to an optional weighting on the level of whole training utterances, as can be seen in the following derivation of the iteration equations for the case of Gaussian mixture densities.

An optimization of the class of discriminative training criteria defined above 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. Thus, these criteria optimize the class separability according to the words under consideration of the language model.

2.1. Parameter Optimization

One possibility to maximize discriminative training criteria consists of a gradient descent with the following reestimation formula for the parameters:

$$\hat{\lambda} = \lambda + \epsilon \cdot \frac{\partial F_D(\lambda)}{\partial \lambda}.$$

A mixture density for an acoustic observation vector x given an HMM state s shall be denoted by $p(x|s,\lambda_s)$. The according parameters λ_s of a mixture density are the weights c_{sl} and parameters λ_{sl} of densities l of the mixture, and mixture densities shall be calculated in maximum approximation. Then the derivative of the general discriminative criterion F_D with respect to parameters θ_{sl} is given by:

$$\frac{\partial F_D(\lambda)}{\partial \theta_{sl}} = \Gamma_{sl}\left(\frac{\partial \log c_{sl} p(x|\lambda_{sl})}{\partial \theta_{sl}}\right),$$

where the discriminative averages Γ_{sl} are defined by:

$$\begin{aligned}\Gamma_{sl}\left(g(x)\right) &= \alpha \sum_{r=1}^{R} f_r \sum_{t=1}^{T_r} \left(\gamma_{rt}(s;W_r) - \gamma_{rt}(s)\right) \\ &\qquad \cdot \eta_{rt}(l|s)\; g(x_{rt}), \\ \Gamma_{s}\left(g(x)\right) &= \sum_{l} \Gamma_{sl}\left(g(x)\right),\end{aligned} \tag{1}$$

where we have utterance weights f_r which have to be considered if the smoothing function f is not the identity,

$$f_r = f'\big(\log \frac{P_\lambda(X_r|W_r)P(W_r)}{\sum_W P_\lambda(X_r|W)P(W)}\big).$$

Applying the maximum approximation for the calculation of mixture densities, the density probabilities $\eta_{rt}(l|s)$ are determined by

$$\eta_{rt}(l|s) = \delta\left(l, \text{argmax}_k\; c_{sk} p(x_{rt}|\lambda_{sk})\right),$$

with the Kronecker delta $\delta(i,j)$. The discriminative averages also make use of the *Forward-Backward* (FB) probabilities of the spoken word sequence W_r:

$$\gamma_{rt}(s;W_r) \;=\; p_\lambda(s_t = s|X_r, W_r),$$

and the generalized FB probabilities for the total of all competing word sequences W defined by the sets $\mathcal{M}_r$:

$$\gamma_{rt}(s) = \sum_{W\in\mathcal{M}_r} \frac{p^\alpha(X_r,W)}{\sum_{V\in\mathcal{M}_r} p^\alpha(X_r,V)}\; \gamma_{rt}(s;W).$$

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

Using the Viterbi approximation [4], i.e. calculating the FB probabilities from the according time alignment, the sum over all competing word hypotheses for calculation of the generalized FB probability could be separated from the time alignment. Then the according word-posterior probabilities needed could be calculated efficiently by applying a FB calculation scheme on the basis of word hypotheses on a word graph. Thus word graphs could also be used, if α is not 1, which would not be possible if the word graph FB scheme is applied on state level already, as done for the MMI criterion in [10]. It should be noted that the calculation of word-posterior probabilities also finds applications in other areas of speech recognition like the determination of confidence measures, e.g. [8].

Discriminative training with the MMI criterion usually applies an extended version of *Baum-Welch* training, the EB algorithm [5, 6]. We extended this approach to the general criterion F_D, which could be maximized via the following auxiliary function:

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

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

$$\begin{aligned}\frac{\partial S(\lambda,\hat{\lambda})}{\partial \hat{\lambda}_{sl}} \;=\;& \Gamma_{sl}\left(\frac{\partial \log p(x|s,\hat{\lambda}_s)}{\partial \hat{\lambda}_{sl}}\right) \\ &+ D_s \int dx\; p(x|s,\lambda_s) \frac{\partial \log p(x|s,\hat{\lambda}_s)}{\partial \hat{\lambda}_{sl}}.\end{aligned}$$

Using discriminative averages for writing down reestimation formulae yields expressions which are formally independent of the particular criterion chosen. Thus, differences of criteria are introduced via the discriminative averages only and comparisons could be reduced to this level.

Performing the EB algorithm, we obtain the following reestimation equations for the means μ_{sl}, state specific diagonal variances σ_s^2 and mixture weights c_{sl} of Gaussian mixture densities:

$$\begin{aligned}\hat{\mu}_{sl} &= \frac{\Gamma_{sl}(x) + D_s c_{sl}\mu_{sl}}{\Gamma_{sl}(1) + D_s c_{sl}} \\ \hat{\sigma}_s^2 &= \frac{\Gamma_s(x^2) + D_s(\sigma_s^2 + \sum_l c_{sl}\mu_{sl}^2)}{\Gamma_s(1) + D_s} \\ &\quad - \sum_l \frac{\Gamma_{sl}(1) + D_s c_{sl}}{\Gamma_s(1) + D_s}\; \hat{\mu}_{sl}^2 \\ \hat{c}_{sl} &= \frac{\frac{\Gamma_{sl}^{spk}(1)}{\Gamma_s^{spk}(1)} - \frac{\Gamma_{sl}^{gen}(1)}{\Gamma_s^{gen}(1)} + C_s}{\sum_{l'} c_{sl'}\left[\frac{\Gamma_{sl'}^{spk}(1)}{\Gamma_s^{spk}(1)} - \frac{\Gamma_{sl'}^{gen}(1)}{\Gamma_s^{gen}(1)}\right] + C_s}\; c_{sl}.\end{aligned}$$

An alternative would be to perform gradient descent on the criterion F_D. Doing this and comparing both sets of reestimation formulae we arrive at step sizes for gradient descent [9], which lead to reestimation formulae, which differ only for the variances

by terms containing the squared step sizes of the means of the according mixture:

$$\begin{aligned}\hat{\mu}_{sl,\mathrm{GD}} &= \hat{\mu}_{sl,\mathrm{EB}} \\ \hat{\sigma}^2_{s,\mathrm{GD}} &= \hat{\sigma}^2_{s,\mathrm{EB}} + \sum_l \frac{\Gamma_{sl}(1) + D_s c_{sl}}{\Gamma_s(1) + D_s}(\mu_{sl} - \hat{\mu}_{sl,\mathrm{BW}})^2 \\ \hat{c}_{sl,\mathrm{GD}} &= \hat{c}_{sl,\mathrm{EB}}.\end{aligned}$$

The reestimation formulae for the mixture weights do not result directly from the optimization of the criterion but are smoothed versions for better convergence [5]. For this version the discriminative averages $\Gamma_{s(l)}$, as defined in Equation 1, are separated according to the FB probability for the spoken (spk) word sequence and the generalized (gen) FB probability for the total of all competing word sequences.

Setting $\alpha = 1$ for comparison purposes, we observe only two differences between MCE and MMI. Firstly, for MMI the spoken word sequence is considered for discrimination, whereas it has to be excluded when using MCE. Since the word-posterior probabilities of correct words securely recognized will be nearly 1, the differences of FB probabilities in the discriminative averages for MMI are nearly zero, such that those words do not contribute significantly to reestimation. Secondly, the worse the recognition results for an utterance are, the more it will contribute to MMI reestimation, which is not the case for MCE. For MCE, hopelessly bad recognized utterances together with securely recognized ones are weighted down as a whole according to their posterior probabilities via the smoothing function f.

Fast convergence is achieved if the iteration constants D_s are chosen such that the denominators in the reestimation equations and the according variances are kept positive:

$$D_s = h \cdot \max\left\{D_{s,\min}, \frac{1}{\beta} - \Gamma_s(1)\right\}. \qquad (2)$$

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. Similarly, the iteration parameters C_s for the mixture weights are chosen such that all weights are positive:

$$C_s = \max_l\left\{-\left[\frac{\Gamma^{spk}_{sl}(1)}{\Gamma^{spk}_{s}(1)} - \frac{\Gamma^{gen}_{sl}(1)}{\Gamma^{gen}_{s}(1)}\right], 0\right\} + \epsilon,$$

with a small constant ϵ.

3. RESULTS

Experiments were performed on the *SieTill* corpus [2] for telephone line recorded German continuous digit strings. The *SieTill* corpus consists of approximately 43k spoken digits in 13k sentences for both training and test.

The recognition system for the *SieTill* corpus is based on whole word HMMs using continuous emission distributions. It is characterized as follows:

- Gaussian mixture emission distributions,
- pooled or state dependent variance vectors,
- gender dependent whole word HMMs for 11 German digits including '*zwo*' and gender dependent silence models,
- per gender 132 states plus one for silence,
- 12 cepstral features plus first derivatives and the second derivative of the energy.

Table 1: Comparison of recognition results on the *SieTill* corpus for ML and discriminative training for different acoustic modeling and training techniques.

corp	LDA	var	dns	crit	opt	WER[%]		
						del	ins	tot
test	no	PV	1	ML	-	0.7	1.0	5.6
			4	ML	-	0.4	1.8	4.8
			1	CT	GD	0.8	0.6	3.3
		SV	1	ML	-	0.6	1.6	5.2
			4	ML	-	0.5	1.7	4.6
			25	ML	-	0.4	1.6	4.1
			1	CT	GD	0.8	0.7	3.4
	yes	PV	1	ML	-	0.5	0.7	4.0
				CT	GD	0.5	0.7	2.8
			4	ML	-	0.3	1.0	3.0
				CT	GD	0.6	0.5	2.5
				CT	EB	0.6	0.5	2.6
				WG	GD	0.7	0.5	2.6

The baseline recognizer applies ML training using the Viterbi approximation [4] which serves as a starting point for the additional discriminative training. A detailed description of the baseline system could be found in [11].

In Table 1 the recognition results obtained for several acoustic modeling approaches using maximum likelihood training are indicated by ML. For ML training, state specific variances (SV) gave better results than pooled variances (PV) for both single and mixture densities. The best results for state specific variances were obtained using approximately 25 densities per mixture, whereas for pooled variances the best results were already obtained for approximately 4 densities per mixture. Adding linear discriminant analysis (LDA) to using pooled variances gave our best ML results with 4.0% word error rate for single densities and 3.0% word error rate for mixture densities with approx. 4 densities per mixture. It should be noted that the LDA gave a relative improvement of over 60% in word error rate for pooled variances and still more than 25% compared to state specific variances without LDA.

For CT, an iteration factor of $h = 1.2$ leads to relatively smooth convergence. Fig. 1 shows a plot of the MMI criterion for the male portion of the *SieTill* training corpus for CT using both GD and EB optimization starting from the according ML result using Gaussian mixture densities with approx. 4 densities per mixture, pooled variance vector and LDA. After CT has converged (indicated by the vertical line), a plot of the MMI criterion using word graphs for discrimination (WG) with an average number of about 47 word hypotheses per spoken word is added. Certainly the absolute values of the MMI criterion using CT and WG respectively are not comparable. In a region where CT does not converge any more, the MMI criterion clearly converges, although the word error rate obtained by CT is even slightly better than that for WG (cf. Table 1). The reason for this could be, that an utterance, which is correctly recognized does not contribute to reestimation for CT. Thus, also incorrectly hypothesized word sequences for such utterances are not considered for discrimination using CT, even if

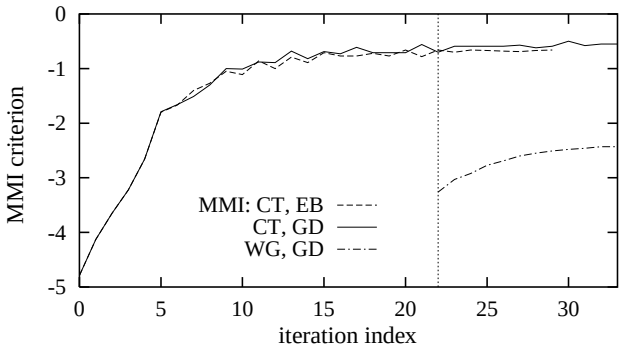

Figure 1: MMI criterion for the male speakers of the *SieTill* training corpus for corrective training (CT) and the use of word graphs (WG).

their posterior probabilities are only marginally smaller than the maximum. Contrarily, using WG would try to reduce the posterior probability of such marginal second best hypotheses, although this might not be necessary, so far as these wrong hypotheses keep being second best at most. Such, this further rearrangement in the posterior probabilites done using WG might have no or even negative effects on the word error rate in comparison to CT, as observed in our experiments. Table 1 summarizes the recognition results for the *SieTill* test corpus using Gaussian emission densities. For different levels of acoustic modeling, we compare ML results with the according discriminative training results using the MMI criterion with corrective training (CT) approximation and gradient descent (GD) optimization. The largest improvements using CT were obtained for our simplest system using single densities with pooled variances (PV), where the ML training word error rate was reduced by 40% relatively. Although the initial ML result for single densities with density specific variances (SV) was better than the according result for pooled variances, the improvement obtained by additional CT was smaller, and the word error rate obtained was even slightly smaller than that for CT using single densities with pooled variance. It should be noted, that the result for CT using single densities with density specific variances was even better than the according ML result using mixture densities with 25 densities per mixture. The best results for CT were obtained using mixture densities with 4 densities per mixture, pooled variance (PV) and LDA, leading to a word error rate of 2.5%, which is the same as reported in [2]. Still, the according result for CT with single densities is slightly better than the results for ML training using 4 densities per mixture. Finally, the best CT result using GD optimization was compared to CT using EB optimization, showing no significant difference for mixture densities as was the case for single densities [9].

4. CONCLUSION

We presented a formally unifying approach for a class of discriminative training criteria and optimization methods including the *Maximum Mutual Information* (MMI) and the *Minimum Classification Error* (MCE) criterion which were compared. For the MMI criterion, experiments were performed on the *SieTill* corpus. Relative improvements in word error rate of up to 40% compared to ML training were obtained, and MMI training using single densities always produced better results than ML training using mixture densities. For the best initial result using ML training, the relative improvement obtained by a subsequent MMI training was about 1/6, leading to a word error rate of 2.5%. This result, which was obtained for corrective training, i.e. using only the best recognized word sequence in addition to the spoken word sequence, could not be improved by using the word graph based discriminative training.

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

5. REFERENCES

[1] 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*, Minneapolis, MN, Vol. 2, pp. 652-655, April 1993.

[2] 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*, Philadelphia, PA, Vol. I, pp. 252-255, October 1996.

[3] R. Haeb-Umbach, H. Ney. "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition," Proc. *Int. Conf. Acoustics, Speech and Signal Process. 1992*, San Francisco, CA, Vol. 1, pp. 13-16, March 1992.

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

[5] 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.

[6] 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.

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

[8] T. Kemp, T. Schaaf. "Estimating Confidence using Word Lattices," Proc. *1997 Europ. Conf. on Speech Communication and Technology*, Rhodes, Greece, Vol. 2, pp. 827-830, September 1997.

[9] R. Schlüter, W. Macherey, S. Kanthak, H. Ney, L. Welling. "Comparison of Optimization Methods for Discriminative Training Criteria," Proc. *1997 Europ. Conf. on Speech Communication and Technology*, Rhodes, Greece, Vol. 1, pp. 15-18, September 1997.

[10] V. Valtchev, J. J. Odell, P. C. Woodland, S. J. Young. "Lattice-Based Discriminative Training For Large Vocabulary Speech Recognition," In Proc. *Int. Conf. Acoustics, Speech and Signal Process. 1996*, Atlanta, GA, Vol. 2, pp. 605-608, May 1996.

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