Investigations on Error Minimizing Training Criteria for Discriminative Training in Automatic Speech Recognition

Wolfgang Macherey, Lars Haferkamp, Ralf Schlüter, Hermann Ney

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen University, 52056 Aachen, Germany
{w.macherey,haferkamp,schluter,ney}@informatik.rwth-aachen.de

Abstract

Discriminative training criteria have been shown to consistently outperform maximum likelihood trained speech recognition systems. In this paper we employ the *Minimum Classification Error* (MCE) criterion to optimize the parameters of the acoustic model of a large scale speech recognition system. The statistics for both the correct and the competing model are solely collected on word lattices without the use of N-best lists. Thus, particularly for long utterances, the number of sentence alternatives taken into account is significantly larger compared to N-best lists. The MCE criterion is embedded in an extended unifying approach for a class of discriminative training criteria which allows for direct comparison of the performance gain obtained with the improvements of other commonly used criteria such as *Maximum Mutual Information* (MMI) and *Minimum Word Error* (MWE). Experiments conducted on large vocabulary tasks show a consistent performance gain for MCE over MMI. Moreover, the improvements obtained with MCE turn out to be in the same order of magnitude as the performance gains obtained with the MWE criterion.

1. Introduction

Due to improved optimization procedures and increased computational power, discriminative methods have become an important means of estimating the parameters of *Hidden Markov Models* in many state-of-the-art speech recognition systems. Since the first successful application of the *Maximum Mutual Information* (MMI) criterion to large scale speech recognition tasks [1], there has been a growing interest in a class of error minimizing discriminative training criteria, as for example the *Minimum Word Error* (MWE) and the *Minimum Phone Error* (MPE) criterion [2]. In contrast to the MMI criterion, which directly maximizes the posterior probability of the training utterances, MWE and MPE aim at minimizing the expectation of the word and phoneme error rate on training data. The MWE and MPE criterion could be shown to significantly outperform the MMI criterion on many tasks [2, 3].

Another criterion that also ranks among the class of error minimizing criteria is the *Minimum Classification Error* (MCE) criterion, which aims at minimizing a smoothed sentence error on training data [4, 5]. Although the MCE criterion could be shown to give consistently better results on small vocabulary tasks compared to the MMI criterion [6, 7, 11], there are only few publications that investigate the use of MCE on large vocabulary tasks [8, 9]. One reason is that the MCE criterion requires the exclusion of the correct class from the set of all competing classes. In automatic speech recognition this means that the spoken word sequence has to be removed from the set of all possible word sequences. However, this might be difficult if the set of competing word sequences is encoded as a word lattice: since a lattice may contain multiple alignments and pronunciation variants of the spoken utterance, its constituting arcs may not uniquely be assigned to the correct or a competing sentence without changing the structure of the lattice. One possible remedy is the use of N-best lists, which was examined e.g. in [10]. Another alternative is the use of finite state machines. Here, for each utterance a corresponding transducer is to be built that encodes the set of competing word sequences considered for discrimination. The spoken word sequence can then be excluded using standard operations. However, in general this is less efficient than directly using a word lattice, since in the worst case the exclusion of a string from a transducer may lead to N-best lists.

In this paper we propose a new algorithm that directly operates on word lattices without changing the lattice structure. The use of word lattices for estimating the parameters of the acoustic model under the MCE criterion was first presented in [11]. Though this work already contained the basic principles of the algorithm presented here, it still required N-best lists in order to find all sentence hypotheses in the word graph that correspond with the spoken word sequence. In this paper, the statistics that are necessary in order to train the acoustic model parameters under the MCE criterion are solely extracted from word graphs without using information derived from N-best lists. Experiments conducted on various settings of the *Wall Street Journal* tasks show significant performance gains of the MCE criterion over the MMI criterion. However, while this outcome might have been expected based on the results of experiments conducted on small vocabulary tasks [7, 11], it is much more surprising that, in all settings, the improvements gained by the MCE criterion are always in the same order of magnitude as the improvements obtained with the MWE criterion.

2. An Extended Unifying Approach for a Class of Discriminative Training Criteria

In [6] a unifying view for a class of discriminative training criteria was presented that allowed for directly comparing the performance gains obtained with the MMI and MCE criterion. In this section this approach will be extended such that it also comprises the MWE and MPE criterion. Let $r = 1, ..., R$ denote the training utterances, each consisting of a sequence X_r of acoustic observation vectors $x_{r1}, ..., x_{rT_r}$ and the corresponding word sequence $W_r = w_{r1}, ..., w_{rN_r}$. The emission and the language model probability are denoted by $p_\theta(X_r|W_r)$ and $p(W_r)$. The language model probabilities are supposed to be given. Hence the parameter θ comprises the set of all parameters of the acoustic model. Finally, let $\mathcal{M}_r$ denote a set

Table 1: A class of discriminative training criteria contained in the extended unifying approach.

criterion	smoothing function $f(z)$	alternative word sequences $\mathcal{M}_r$	exponent α	gain function $\mathcal{G}(W, W_r)$
Maximum Likelihood	z	$\emptyset$	-	$\delta(W, W_r)$
Maximum Mutual Information	z	all (recognized)	1	
Corrective Training		best (recognized)	∞	
Minimum Classification Error	$-\dfrac{1}{1+e^{2\varrho z}}$	all without W_r	*free*	
Falsifying Training		best (recognized) $\neq W_r$	∞	
Diversity Index	$-\frac{1}{\varrho}(1-e^{\varrho z})$	all (recognized)	*free*	
Jeffreys	$-\frac{z}{1-z}$	all (recognized)	1	
Minimum Word/Phone Error	$\exp(z)$	all (recognized)	1	$\mathcal{A}(W, W_r)$

of word sequences which are considered for discrimination in utterance r. A class of discriminative training criteria $\mathcal{F}$ can then be defined by:

$$\mathcal{F}\big(\theta; f, \alpha, \mathcal{G}, \{\mathcal{M}_r\}\big) = \tag{1}$$

$$\frac{1}{R}\sum_{r=1}^{R} f\left(\log\left[\frac{\sum_W p_\theta^\alpha(X_r|W)\cdot p^\alpha(W)\cdot \mathcal{G}(W,W_r)}{\sum_{W\in\mathcal{M}_r} p_\theta^\alpha(X_r|W)\cdot p^\alpha(W)}\right]^{1/\alpha}\right)$$

The choice of the set of alternative word sequences together with the optional smoothing function f, the weighting exponent α, and the gain function $\mathcal{G}$ determine the particular criterion. Table 1 lists some of the criteria included in this approach. All criteria except MWE and MPE discriminate the spoken word sequence, which is expressed by the choice of the *Kronecker* function δ for the gain function. In contrast to this, the numerator in the MWE criterion considers the sum over *all* possible word sequences weighted with a measure for accuracy $\mathcal{A}$. Note that all criteria are to be maximized, which cause the negative sign in the smoothing function of the MCE criterion, its maximum approximation, the *Falsifying Training*, the *Diversity Index*, and the *Jeffreys* criterion. The derivative of the unified criterion wrt. to the parameter set θ yields the well known re-estimation equations, which can be found e.g. in [6] for the MMI and MCE criterion, and in [2] for the MWE an MPE criterion.

For the special case $\varrho = 1/2$ and $\alpha = 1$ the MCE criterion directly minimizes the expectation of the sentence error, i.e. the sum over $1 - p(W_r|X_r)$ for all training utterances r. Smaller values of ϱ smooth the sentence error, and thus, the criterion minimizes an approximated error rate. This often improves robustness towards outliers in the training data [11]. The same holds for the *Diversity Index*, which, in the case of setting $\varrho = 1$, is equivalent to the *Gini* criterion. In contrast to both the *Diversity Index* and the MCE criterion, the MWE and MPE criterion minimize the expectation of an unsmoothed error rate. A possible extension would therefore be to integrate a smoothing term into the gain function of the MWE/MPE criterion, which might help to further reduce the error rate on unseen data.

3. MCE on Word Lattices

Using the MCE criterion, the set of competing hypotheses comprises all word sequences W that are represented in a word graph, except the spoken sequence W_r. In order to determine the word probabilities on word graphs similar to the MMI criterion, the spoken word sequence has to be excluded from the word graph. However, in general removing a sentence hypothesis from a word graph would change its structure and would result in an increased lattice size, because particular words of the spoken word sequence might be part of other sequences, too. Therefore, the sum over all word sequences in the word graph (represented by $\mathcal{M}_r$) including the spoken word sequence is performed first, which afterwards is subtracted from the probability of the spoken word sequence. Thus the probability q of hypothesizing a word w within the time frames $[t_b, t_e]$ under the MCE criterion can be written as [7, 11]:

$$q_{[t_b,t_e]}(w|X_r) = \frac{\sum\limits_{\substack{\{W\in\mathcal{M}_r|W\neq W_r \\ \wedge w_{[t_b,t_e]}\in W\}}} p_\lambda^\alpha(X_r, W)}{\sum\limits_{\{V\in\mathcal{M}_r|V\neq W_r\}} p_\lambda^\alpha(X_r, V)}$$

$$= \frac{\sum\limits_{\substack{\{W\in\mathcal{M}_r| \\ w_{[t_b,t_e]}\in W\}}} p_\lambda^\alpha(X_r, W) - \sum\limits_{\substack{\{W\in\mathcal{M}_r|W=W_r \\ \wedge w_{[t_b,t_e]}\in W\}}} p_\lambda^\alpha(X_r, W)}{\sum\limits_{V\in\mathcal{M}_r} p_\lambda^\alpha(X_r, V) - \sum\limits_{\{V\in\mathcal{M}_r|V=W_r\}} p_\lambda^\alpha(X_r, V)} \tag{2}$$

Besides the best time alignment of the spoken word sequence, a word graph may contain further copies of the spoken sequence that may vary in boundary times and pronunciation variants. Typically, the scores of these copies differ only marginally from the score of the best alignment. Hence, for MCE training it is necessary to detect and label *all* alignments of the spoken word sequence occurring in the word graph so that the sum over the joint probabilities of these sentence hypotheses can be subtracted afterwards from the word probabilities (cf. Eq. (2)).

4. Experimental Results

Experiments were conducted on three settings of the *Wall Street Journal* (WSJ) corpora [12, 13]. The three tasks differ in the amount of training data and the vocabulary sizes. Table 2 summarizes some corpus statistics.

The WSJ0 recognition system uses 2000 decision-tree based gender independent within-word triphone states plus one state for silence. The states are assigned to Gaussian mixture distributions with a total of 149k densities sharing one common diagonal variance vector. The observation vectors consist of 16 cepstral features together with the first derivatives and the second derivative of the energy. Each five adjacent input frames are concatenated (including derivatives: $5 \times 33 = 165$ input features) and reduced to 33 output features via a linear discriminant analysis (LDA). The baseline recognizer applies *Maximum Likelihood* (ML) training using the Viterbi approximation and

Table 2: Corpus statistics and vocabulary sizes on the *Wall Street Journal* (WSJ0) task and the *North American Business* (NAB) corpus.

corpus	WSJ0			NAB-20k / NAB-65k		
	train	dev	eval	train	dev	eval
acoustic data [h]	15:17	0:46	0:40	81:23	0:48	0:53
# speakers	84	10	8	284	20	20
# sentences	7240	410	330	37474	310	316
# running words	130976	6784	5353	642074	7387	8193
# lexicon words	10133	5007		15013	64735	

achieves a *word error rate* (WER) of 4.14% on the combined development[1] plus the evaluation set (cf. Tab. 4).

The Nov. '94 *North American Business* (NAB) training corpus consists of the 84 speakers of the WSJ0 corpus plus 200 additional speakers from the WSJ1 corpus. Tests were performed on the NAB Nov. '94 *Hub-1* development and evaluation corpus. Both the 20k and the 65k recognition system use 7000 decision-tree based gender independent across-word triphone states plus one state for silence. The system employs Gaussian mixture distributions with a total of 412k densities and one globally pooled diagonal variance vector. As in the WSJ0 setting, 16 cepstral features together with their first derivatives and the second derivative of the energy are used. Each three consecutive observation vectors are concatenated and projected onto a 32 dimensional feature vector via a LDA. The ML trained recognizer achieves a WER of 11.47% for the 20k system and 9.28% for the 65k system on the combined development and evaluation corpus (cf. Table 4).

In all discriminative experiments, the ML trained system was used to generate high density word lattices for both the numerator and the denominator model. The numerator lattices were merged into the denominator lattices at which hypotheses that were newly added to the denominator lattice or that matched a denominator hypothesis were tagged with the label "correct" in order to identify them for the MCE training. To reduce the computational costs during discriminative training, the lattice sizes were reduced via a forward-backward pruning. The resulting word graph densities are shown for the WSJ0 corpus in Table 3. For all iterations of the discriminative training the hypotheses encoded in the word lattices were re-aligned within their boundary times (the Viterbi segmentation points) as determined in the initial recognition phase. For MCE training the smoothing constant ϱ was set to 0.04. Since MWE is reported to give slightly better results than MPE on the WSJ tasks [3], we used the MWE criterion for the comparison.

Figure 1 depicts the evolution of the WER on the combined development and evaluation set of the WSJ0 corpus in the course of the iteration process for the MMI, the MCE, and the MWE criterion. The relatively large number of training iterations that were necessary in order to find the best parameter set (wrt. test set performance) is contrary to what is reported in literature. Usually, it requires 4-8 iterations only before discriminative training starts to overfit the training data and, hence, deteriorates test set performance [2]. However, in this setting, the effect is caused by using a pooled variance vector. Since in discriminative training convergence speed is usually adjusted under a positive variance constraint, using state or density specific variances provides much more constraints in order to chose the "correct" step size, which often results in faster convergence [14].

[1] Since the official WSJ0 corpus does not provide a development set, the 410 sentences were extracted from 10 new speakers of the *North American Business* task.

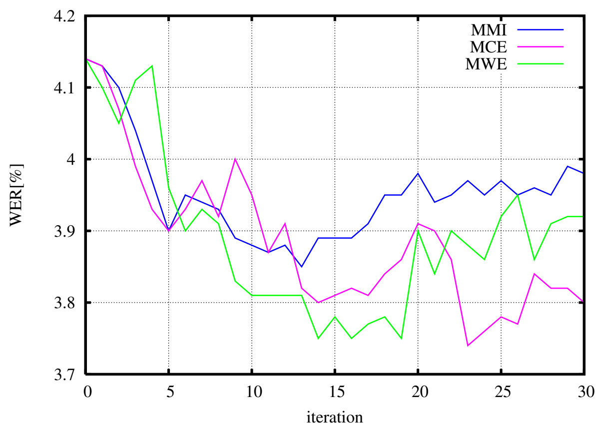

Figure 1: Evolution of the word error rate (WER) on the combined development plus evaluation set of the WSJ0 corpus in the course of the iteration process for the *Maximum Mututal Information* (MMI), *Minimum Classification Error* (MCE), and *Minimum Word Error* (MWE) criterion.

Table 4 shows the word error rates for the discriminative criteria MMI, MCE, and MWE. Compared with the ML trained system, the MMI criterion results in a word error rate of 3.85%, which is a relative improvement of 7%. Compared to this, the MCE and the MWE criterion result in 3.75% WER, which is a relative improvement of 10% compared to the ML baseline, and ca. 4% performance gain in comparison with the MMI result. Surprisingly the MWE criterion does not seem to be superior to the MCE criterion for this setting. Though both criteria lead to a consistent reduction in terms of WER compared to the MMI criterion, the performance gain might be less caused by a direct minimization of a loss function but by the stronger robustness of MCE and MWE towards outliers in the training data. Moreover, MWE directly minimizes the expectation of the word error, although the decision in recognition is made on a sentence level. Thus, MWE would potentially outperform MCE in combination with *Bayes Risk* minimizing decision rules [15].

The experiments conducted on the NAB-20k and NAB-65k tasks (cf. Table 4) show similar results. Though the absolute improvement of MCE compared with MMI is only marginal, MCE gives consistently good results on both the development and the evaluation set. As in the experiments conducted on the WSJ0 corpus, the smoothing constant was set to 0.04. Note that the NAB-65k system uses the same acoustic models as the NAB-20k system, yet with an extended pronunciation lexicon that reduces the number of unknown words on test data from 2.7% to 0.7%. In both settings the MWE criterion leads to slightly yet not significant increases in terms of WER compared to MMI. The improvements observed range in the same order of magnitude as the performance gains reported in [3, p. 114].

Table 3: Word graph densities on the training, development, and evaluation set of the WSJ0 corpus together with the respective *graph error rates* (GER).

corpus	WSJ0			
	train	dev	eval	dev+eval
avg. #arcs per spk. word	210.86	237.32	261.89	248.28
avg. #arcs per rec. word	159.84	189.13	202.64	195.15
avg. #arcs per frame	99.89	81.07	78.42	79.83
GER [%]	0.0	0.22	0.09	0.16

Table 4: Word error rates (WER) and sentence error rates (SER) on the *Wall Street Journal* (WSJ0) corpus and the *North American Business* (NAB) corpora for a class of discriminative training criteria, including the *Maximum Mutual Information* (MMI) criterion, *Minimum Classification Error* (MCE), and *Minimum Word Error* (MWE).

corpus	WSJ0					NAB-20k					NAB-65k				
	dev		eval		dev+eval	dev		eval		dev+eval	dev		eval		dev+eval
	WER	SER	WER	SER	WER	WER	SER	WER	SER	WER	WER	SER	WER	SER	WER
ML	4.48	40.2	3.72	34.9	4.14	11.48	73.9	11.46	76.3	11.47	9.21	67.1	9.35	71.2	9.28
MMI	4.16	38.8	3.46	33.3	3.85	11.18	73.6	11.02	74.1	11.10	8.93	67.7	8.97	69.0	8.95
MCE	3.98	37.1	3.44	33.0	**3.74**	11.11	74.2	10.97	75.3	**11.04**	8.81	67.1	9.04	69.3	**8.93**
MWE	4.05	37.3	3.36	31.2	3.75	11.17	73.2	11.22	75.3	11.19	8.84	67.7	9.11	70.6	8.98

5. Conclusions

In this paper we investigated the use of the *Minimum Classification Error* (MCE) criterion for training the acoustic model parameters of a large scale speech recognition system. In contrast to other studies, all statistics necessary for re-estimating the model parameters under the MCE criterion have been determined on word lattices for both the correct and the competing model. Thus, particularly for long utterances, the number of sentence alternatives taken into account in training is significantly larger compared to N-best lists.

The investigations were carried out within an extended unifying framework for discriminative training criteria that, besides MCE, also includes the *Maximum Mutual Information* (MMI) and the *Minimum Word Error* (MWE) criterion. While MCE showed consistently better results compared to MMI of up to 4% relative on the *Wall Street Journal* task, its performance in terms of *word error rate* (WER) turned out to be in the same order of magnitude as MWE. Compared to a *Maximum Likelihood* trained system, the MCE criterion lead to relative improvements of between 4% and 10% in terms of WER.

6. Acknowledgements

This work was partially funded by the European Union under the integrated project "TC-STAR - Technology and Corpora for Speech to Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org)".

7. References

[1] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 25–48, 2002.

[2] D. Povey and P. C. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, 2002*, vol. 1, Orlando, FL, May 2002, pp. 105–108.

[3] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Dept. of Eng., Cambridge Univ., Cambridge, August 2004.

[4] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," in *IEEE Transactions on Signal Processing*, vol. 40, no. 12, December 1992, pp. 3043–3054.

[5] W. Chou, C.-H. Lee, and B.-H. Juang, "Minimum Error Rate Training based on N-Best String Models," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, Minneapolis, MN, USA, April 1993, pp. 652–655.

[6] R. Schlüter and W. Macherey, "Comparison of discriminative training criteria," in *1998 Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, Seattle, WA, May 1998, pp. 493–496.

[7] R. Schlüter and W. Macherey and B. Müller and H. Ney, "Comparison of discriminative training criteria and optimization methods for speech recognition," *Speech Communication*, vol. 34, no. 1, pp. 287–310, May 2001.

[8] E. McDermott and T. J. Hazen, "Minimum classification error training of landmark models for real-time continuous speech recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, Montreal, Canada, May 2004, pp. 937–940.

[9] E. McDermott and S. Katagiri, "Minimum classification error for large scale speech recognition tasks using weighted finite state transducers," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, Philadelphia, PA, March 2005, pp. 113–116.

[10] K. K. Paliwal, M. Bacchiani, and Y. Sagisaka, "Minimum classification error training algorithm for feature extractor and pattern classifier in speech recognition," in *1995 Europ. Conf. on Speech Communication and Technology*, vol. 1, Madrid, Spain, September 1995, pp. 541–544.

[11] W. Macherey, "Implementation and comparison of discriminative training methods for automatic speech recognition," Diploma Thesis, Lehrstuhl für Informatik VI, RWTH Aachen University, Aachen, November 1998.

[12] D. S. Pallett, J. G. Fiscus, W. M. Fisher, J. S. Garofolo, B. A. Lund, and M. A. Przybocki, "1994 Benchmark test for the ARPA spoken language program," in *ARPA Human Language Technology Workshop*, Austin, TX, January 1995, pp. 5–36.

[13] F. Kubala, "Design of the 1994 CSR benchmark tests," in *ARPA Human Language Technology Workshop*, Austin, TX, January 1995, pp. 41–46.

[14] W. Macherey, R. Schlüter, and H. Ney, "Discriminative training with tied covariance matrices," in *8th Int. Conf. on Spoken Language Processing*, vol. 1, Jeju Island, Korea, October 2004, pp. 681–684.

[15] V. Goel, S. Kumar, and W. Byrne, "Segmental minimum Bayes-risk decoding for automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 3, pp. 234–249, May 2004.