Use of Incrementally Regulated Discriminative Margins in MCE Training for Speech Recognition

Dong Yu, Li Deng, Xiaodong He, and Alex Acero

Microsoft Research
One Microsoft Way, Redmond, WA 98052, USA
{dongyu, deng, xiaohe, alexac}@microsoft.com

Abstract

In this paper, we report our recent development of a novel discriminative learning technique which embeds the concept of discriminative margin into the well established minimum classification error (MCE) method. The idea is to impose an incrementally adjusted "margin" in the loss function of MCE algorithm so that not only error rates are minimized but also discrimination "robustness" between training and test sets is maintained. Experimental evaluation shows that the use of the margin improves a state-of-the-art MCE method by reducing 17% digit errors and 19% string errors in the TIDigits recognition task. The string error rate of 0.55% and digit error rate of 0.19% we have obtained are the best-ever results reported on this task in the literature.

Index Terms: discriminative training, margin, minimum error

1. Introduction

Discriminative training has been a prominent theme in recent speech recognition research and system development; e.g., [2][3][5][6][10][12][13][14][15]. The essence of these discriminative training algorithms is the adoption of the cost functions that are directly or indirectly related to the empirical error rate in the training data. These cost functions serve as the objective functions for optimization, and the related empirical error rate may be either at the sentence string level [6][10], at the super-string level [3], at the sub-string level (i.e., word or phone in a sentence) [13], or at the isolated word/phone token level [10][14].

One key insight from modern machine learning research (e.g., [9][1][16]) is that when the empirical training error rate is optimized by a classifier or recognizer, only a biased estimate of the true error rate is obtained. How large the bias is depends on the complexity of the recognizer and the task (as quantified by the VC dimension). The analysis and experimental results reported in this paper show that this bias can be quite substantial even for a simple HMM recognizer applied to a simple digit recognition task. Another key insight from machine learning research suggests that one effective way to reduce the above bias for improving generalization performance is to increase "margins" in the training data; i.e., making the correct samples classified well away from the decision boundary. It is desirable to use such large margins for achieving lower test errors, even if this may result in higher empirical errors in training.

Prior to the work presented in this paper, most discriminative learning techniques in speech recognition research have focused on the issue of empirical error rates, and not on the issue of "margins" and related generalization. One notable exception is the recent work of [5], where margins defined from the Gaussian HMM for positive samples are optimized with constraints, while throwing away negative samples. Standard MCE (minimum classification error) training [6] was carried out before margin optimization, reducing the impact of negative examples. While the technique of [5] may not generalize to more difficult tasks than had been evaluated, the positive experimental results published in [5] and the related insights suggest a fruitful direction to explore the effect of large margins in the overall speech recognition generalization performance.

In this paper, we present an alternative technique to [5] for incorporating margins in the discriminative training. In contrast to [5] where the MCE-trained HMMs are used as the initial model for the subsequent large-margin HMM training with positive samples only (i.e., throwing away all negative samples), we integrate both error rate minimization and margin enhancement into a single framework. Significant performance improvement over [5] has been achieved using the identical recognition task.

The rest of the paper is organized as follows. In section 2, we introduce our novel technique that incorporates the discriminative margin in a generalized version of the MCE training. Detailed experimental evaluation of this technique is presented in Section 3. Finally, we summarize and conclude the paper in section 4, with a discussion on the ongoing work to further validate the effectiveness of the new training technique.

2. MCE Incorporating Discriminative Margin

The integrated technique for both error rate minimization and margin enhancement presented in this paper is implemented in the MCE framework with modification. In this section, we will review the conventional MCE framework first, and then describe the modification.

2.1. MCE technique and its implementation

Conventional MCE learning [6][10][14] minimizes the smoothed sentence or string-level error rate. We use r=1,…,R as the index for "token" or "string" (e.g., a single sentence or utterance) in the training data, and each token consists of a "string" of a vector-valued observation data sequence: $X_r=x_{r,1}$, …, $x_{r,Tr}$, with the corresponding label (e.g., word) sequence: S_r = $w_{r,1},\ldots, w_{r,Nr}$. That is, S_r denotes correct label sequence for token r. Further, we use s_r to denote all possible label sequences for the r-th token, including the correct label sequence S_r and all other incorrect label sequences.

In MCE, a loss function for a single utterance X_r is defined. The loss function has the desirable property that it is close to zero if the string is correctly recognized and close to one if it is incorrectly recognized. The most popular smooth

function that gives this property is the following sigmoid function.

$$l_r\left(d_r(X_r,\Lambda)\right)=\frac{1}{1+e^{-\alpha d_r(X_r,\Lambda)}}, \quad (1)$$

where $d_r(X_r,\Lambda)$ is called the misclassification measure and Λ is the model parameters to be trained. For the popular one-best MCE training when only top one incorrectly recognized string is used as the "competitive candidate" for discriminative training, $d_r(X_r, \Lambda)$ is the log-likelihood distance between the correct string, $S_{r,c}$, and the incorrect or competitive string, denoted as $S_{r,e}$, i.e.,

$$d_r(X_r,\Lambda)= - \log p_\Lambda(X_r,S_{r,c})+\log p_\Lambda(X_r,S_{r,e}). \quad (2)$$

Substituting (2) into (1) gives

$$l_r\left(d_r(X_r,\Lambda)\right)=\frac{p^\alpha{}_\Lambda(X_r,S_{r,e})}{p^\alpha{}_\Lambda(X_r,S_{r,e}|\Lambda)+p^\alpha{}_\Lambda(X_r,S_{r,c}|\Lambda)} \quad (3)$$

For the more general N-best MCE training where top N>1 (instead of only one) incorrectly recognized string is used as the "competitive candidates", a soft-max function has been widely used. In our implementation, we have approximated the soft-max by a simpler form. This leads to the loss function for the N-best version of MCE of

$$l_r\left(d_r(X_r,\Lambda)\right)=\frac{\sum_{s_r,s_r\neq S_r} w_{MCE}(s_r)p^\alpha_\Lambda(X_r,s_r)}{\sum_{s_r} w_{MCE}(s_r)p^\alpha_\Lambda(X_r,s_r)} \quad (4)$$

where $w_{MCE}(s_r)$ is a weighting factor of $s_r \neq S_{r,c}$, and we assign $w_{MCE}(S_{r,c}) \equiv 1$. Next, the loss function at the string level is defined to be the sum of the loss functions of individual string tokens:

$$L_{MCE}(\Lambda)=\sum_{r=1}^{R} l_r\left(d_r(X_r,\Lambda)\right) \quad (5)$$

Now, minimizing the string level loss function of $L_{MCE}(\Lambda)$ in (5) is equivalent to maximizing the MCE objective function (R is the total number of training sentences):

$$O_{MCE}(\Lambda)=R-L_{MCE}(\Lambda)=\sum_{r=1}^{R}\frac{p^\alpha_\Lambda(X_r,S_{r,c}|\Lambda)}{\sum_{s_r} w_{MCE}(s_r)p^\alpha_\Lambda(X_r,s_r)} \quad (6)$$

We have implemented the MCE algorithm that maximizes (6) not by gradient ascend (as in generalized probabilistic descent or GPD [6]) but by a special technique of optimization via growth transformation. This implementation is an improved version upon that as originally proposed in [3]. The improvement lies in converting the super-string-level objective function in [3] into a normal string-level objective function for MCE. This conversion is accomplished via a non-trivial mathematical framework (see details in [4]), which results in a rational function that is then subject to optimization by growth transformation or extended Baum-Welch algorithm. We found that in our growth transformation based optimization, much fewer iterations are required for empirical convergence than those typically required by the gradient based GPD [6][14].

2.2. Incorporating discriminative margin in MCE

Given a fixed classifier or recognizer which defines decision boundaries for all possible pairs of classes, a "margin" is defined for each training token as the difference between the score of this token by the correct class and that by the most competitive class. A positive difference gives a positive sample, and a negative one gives a negative sample. A large (positive) margin implies a wide tolerance gap. A recognizer with a larger margin in magnitude gives more robust discrimination than that with a smaller one, but it may not give lower empirical error rates in the training data especially for multi-class tasks such as in speech recognition.

The concept of margin interpreted as the tolerance gap can be readily incorporated into MCE by using a negative (incrementally adjusted or iteration (I)-dependent) parameter $\beta(I)<0$ in the more general definition of the loss function:

$$l_r\left(d_r(X_r,\Lambda)\right)=\frac{1}{1+e^{-\alpha d_r(X_r,\Lambda)+\beta(I)}} \quad (7)$$

In the conventional MCE, β has been invariably set to be zero (e.g., [6][3][10][14]), which gives (1). And since the margin provided by $\beta(I)$ is a component of the loss function in (7) that determines the empirical discrimination power, we call it "discriminative margin".

In (7), the "iteration" argument I in $\beta(I)$ signifies that the actual value of β at iteration I is regulated by incremental adjustment from a smaller (in magnitude) negative value to a larger one. Small-magnitude negative values of β in early iterations provide low margins while not sacrificing significant reduction in empirical errors in training. Once the error pattern becomes adjusted to the new one at the iteration, an increment of β from $\beta(I)$ to $\beta(I+1)$ at the next iteration will have a similarly small effect on the empirical errors while achieving relatively larger margins that help reduce test errors. In addition, the incrementally adjusted margins help bring incorrectly classified training tokens that are far away from the center of the sigmoid function across the center faster than without using such margins. This is because the slopes of the sigmoid corresponding to these tokens are small and thus would be moved slowly by MCE without incremental margins.

We now use Fig. 1 with a two-class special case to illustrate the use of discriminative margins in MCE. Tokens shaped as circles in Fig. 1 are from class 1 and those as triangles are from class 2. Without a margin (upper two sub-figures for class 1 and class 2, respectively), as in the conventional MCE, the circle token near *d=0* for class 1 will contribute to model adjustment since it incurs some loss and it is near the decision boundary where the slope of the sigmoid is large. But after model adjustment which moves that token to the left, the slope of the sigmoid becomes much smaller and hence model adjustment soon stops. (The same process applies to the triangle token near *d=0* for class 2).

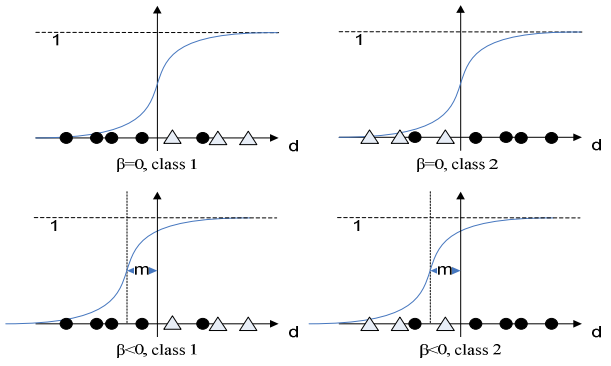

Figure 1. Illustration of the use of discriminative margins in MCE and its desirable effects for a two-class case.

After the margin is introduced (lower two sub-figures for

class 1 and class 2, respectively, in Fig. 1) by shifting the sigmoid function to the left with the magnitude of *m*, the circle token and the triangle token for class 2 (both near *d=0)* tend to move to the left over a distance at least *m* units longer than in the earlier case. Further, when the shift of the sigmoid function is done incrementally, a greater final distance or discriminative margin can be achieved.

3. Experimental Results

We have evaluated our technique described in Section 2.2 above using the TIDIGITS corpus [8], in a standard experimental setup consistent with the prior work on this task [6][5]. This corpus contains utterances from 326 speakers (111 men, 114 women, and 101 children) collected from 21 regions of the United States. There are a total of eleven words (digits) in the corpus vocabulary (digits of "1" to "9", plus "oh" and "zero"). Each utterance is a connected-digit string, with the number of digits in the string varying between one and seven (except with no six-digit strings). In our experiments, we only use the adult portion of the corpus, which makes up a standard training set of 8623 digit strings (from 55 men and 57 women) and a standard test set of 8700 digit strings (from 56 men and 57 women).

In our experiments, all data are sampled at a rate of 16K Hz. The 33-dimentional acoustic feature vectors are composed of the normalized energy, 10 MFCCs (Mel-Frequency Cepstrum Coefficients) and their first and second order time derivatives. The models used in our experiments are head-body-tail CDHMMs with a different number of Gaussian mixture components for each state. The total number of Gaussian mixture components used in the system is 3284, which is roughly the same as in a nine-state whole word CDHMMs with 32 Gaussian mixtures per state.

The models are trained first using the Maximum Likelihood (ML) criterion. Then, MCE training methods before and after incorporating discriminative margins are applied, both using the ML-trained models to initialize the MCE training. The word error rate (WER) and string error rate (SER) in the test set using the initial ML-trained models are 0.28% and 0.78%, respectively, using tuned insertion penalty of -14.5 and language model weight of -13.25. During the MCE training, α value in (7) is tuned to be 1/120, and all HMM model parameters (except transition probabilities) are updated. This setting gives us the best MCE baseline (i.e., no discriminative margin used or $\beta(I)=0$), with WER of 0.23% and SER of 0.68% (as shown in Table 1). This represents 17.86% relative WER reduction and 12.82% relative SER reduction over the initial ML-trained models.

Table 1: Summary of the Experimental Results

Margin	WER		SER	
	Absolute	Relative reduction	Absolute	Relative reduction
$\beta=0$	0.23%	baseline	0.68%	baseline
Method 1	0.20%	13.04%	0.57%	16.18%
Method 2	0.20%	13.04%	0.57%	16.18%
Method 3	**0.19%**	17.39%	**0.55%**	19.12%

We then train the digit HMMs, also initializing from the ML models, with incrementally regulated discriminative margins $\beta(I)<0$ in the MCE training paradigm. We keep the same α (=1/120) and use three different methods for setting the schedule that regulates $\beta(I)$. The three methods are tested under otherwise identical experimental conditions.

In the first method, $\beta(I)$ is set to a fixed value over the range of [-1, 0] in all iterations. That is, $\beta(I)$ is set to be independent of the iteration number I. We obtain the best result when setting $\beta(I)= -0.8$. Details of the results are plotted in Fig. 2, where the recognition error rates (WER and SER) are shown as a function of the fixed β value over MCE training iterations. These error rates for training and test sets are plotted separately. The initial HMMs for the MCE training with each of the fixed β values are from the ML training. A total of 15 MCE growth-transformation iterations are used for each of the fixed β values.

In the second method, $\beta(I)$ is scheduled to change from neutral (no margin or $\beta =0$) to $\beta = -1$, with a step size of -0.1 during the MCE. That is, $\beta(I)= -0.1*(I-1)$, for I=1,…,11. Fig. 3 shows the WER and SER results (for both training and test sets) as a function of the incrementally reduced $\beta(I)$ value. When $\beta =0$ (rightmost set of the results in Fig. 3), the HMMs are initialized in the MCE training (4 iterations) from the ML-trained models. As $\beta(I)$ becomes incrementally reduced from 1 to 11, the previously MCE-trained models serve as the initial models and additional 4 MCE iterations are used for each new β value.

In the third method, $\beta(I)$ is scheduled to change from +0.4 to -0.5, with a step size of -0.1 also. That is, $\beta(I)= 0.4 - 0.1*(I-1)$, for I=1,…,10. Fig. 4 shows the results in the format similar to Fig. 3, with slightly lower errors.

A close examination of the results of Figs. 2-4 reveals a rather consistent trend about the effects of increasing the discriminative margin on the recognition errors. As the margin enlarges (more negative of β), errors tend to reduce first, and then to reverse the direction as the margin further increases. So the largest margin does not correspond to the lowest error. The figures also show that the lowest training and test error rates do not occur at the same β value. The overall experimental results using the three methods discussed above are summarized in Table 1. In the table, relative error reduction is calculated upon the MCE baseline where the discriminative margin $\beta(I)$ is set to zero. We observe 13.04% relative WER reduction and 16.18% relative SER reduction over the baseline MCE models with Methods 1 and 2. By using Method 3, we have achieved 0.19% absolute WER and 0.55% absolute SER, which translate to 17.39% relative WER and 19.12% relative SER reduction over our MCE baseline. The gain has been tested to be statistically significant.

To the best of our knowledge, the best published results on the TIDIGITS task in the literature [6] have been 0.24% (WER) and 0.72% (SER) using the standard MCE with no margin ($\beta=0$). These are very close to our MCE baseline, with differences likely due to the GPD vs. growth-transformation in the optimization procedure.

4. Summary and Conclusions

Use of large margins to improve robustness and generalization performance of pattern recognition has been well motivated and is a standard practice for discriminative training in machine learning [9][16]. Yet most practices of discriminative training in speech recognition have not embraced the concept of large margins and have been concerned mainly with empirical error rates in the training set [2][3][6][13][14]. The recent work of [5] introduced large

margins in training HMMs for speech recognition, after the HMMs are pre-trained by the standard zero-margin MCE method based on GPD. (We noted recently that this important issue of the generalization ability to test data was also discussed in [7][10] in the context of smoothness of the MCE loss function.) This paper reports an alternative method where margins and empirical errors are jointly optimized in a generalized version of MCE. Superior recognition results on the identical task are obtained by our new method.

The idea behind our new method is the incorporation of incrementally adjusted "margin", over the MCE training iteration, in the loss function of the MCE algorithm. In this way, empirical error rates and the discriminative margins are simultaneously optimized. The tradeoffs of introducing the new "margin" parameter are: 1) increased margins help generalization from the training set to the test set; 2) increased margins also create potential danger of sacrificing discrimination on the training set and of possibly sacrificing discrimination on the test set as well. To strike a balance between these two factors working against each other, we have developed a heuristic technique of incrementally regulating the change of the "margin" parameter over the MCE iterations. Experimental results show that the use of these incrementally changed margins significantly improves the prior art.

We are currently investigating several issues for further validating the effectiveness of the new training method discussed in this paper. First, our MCE training is based on the growth-transformation optimization, more efficient than the conventional gradient-descent GPD optimization. It is not clear whether this difference affects the performance improvement we have observed when incorporating the margins. Second, TIDIGITS is a task with a very low error rate. The strategies that we have developed to balance the empirical error rate and the robustness in generalization performance need to be validated with more complex tasks having a higher error rate. Our preliminary results along these lines have been promising.

5. Acknowledgements

We thank Hui Jiang of York University for useful discussions and the earlier large-margin training method which motivated this work.

6. References

[1] Y. Altun, I. Tsochantaridis, and T. Hofmann, "Hidden Markov Support Vector Machines," Proc. Intern. Conf. Machine Learning, 2003.

[2] L. Deng, J. Wu, J. Droppo, and A. Acero. "Analysis and comparison of two feature extraction/compensation algorithms," IEEE Sig. Proc. Letters, Vol. 12, No. 6, 2005, pp. 477-480.

[3] X. He and W. Chou, "Minimum Classification Error Linear Regression for Acoustic Model Adaptation of Continuous Density HMMs," Proc. ICASSP, 2003.

[4] X. He, L. Deng, and W. Chou, "A novel learning method for hidden Markov models in speech and audio processing," Proc. IEEE Workshop on Multimedia Signal Processing, Oct. 2006, Victoria, BC., in print.

[5] X. Li and H. Jiang, "A constrained joint optimization method for large margin HMM estimation," Proc. ASRU Workshop, 2005.

[6] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," IEEE Trans. Speech Audio Proc., Vol. 5, May 1997.

[7] B.-H. Juang and S. Katagiri, "Discriminative training," ASJ Special Issue, Vol. 3, No. 6, 1992, pp, 333-339.

[8] R.G. Leonard, "A Database for Speaker-independent digit recognition", Proc. ICASSP, 1984.

[9] L. Mason, P. Bartlett, and J. Baxter, "Improved generalization through explicit optimization of margins," Machine Learning, Vol. 38, No. 3, March 2000, pp. 243–255.

[10] E. McDermott. "*Discriminative Training for Speech Recognition*", Ph.D. thesis, Waseda University, 1997.

[11] E. McDermott and S. Katagiri, "Prototype based discriminative training for various speech units," Computer Speech and Language, Vol. 8, 1994, pp. 351-368.

[12] F. Pereira. "Linear models for structure prediction", Proc. Interspeech, Lisbon, 2005, pp. 717-720.

[13] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig. "fMPE: Discriminatively trained features for speech recognition," Proc. DARPA EARS RT-04 Workshop, Nov. 7-10, 2004, Palisades, NY, Paper No. 35, 5 pages.

[14] C. Rathinavalu and L. Deng. "Speech trajectory discrimination using the MCE learning," IEEE Trans. Speech and Audio Proc., Vol.6, 1998, pp. 505-515.

[15] F. Sha and L. Saul, "Large margin Gaussian mixture modeling for phonetic classification and recognition," Proc. ICASSP, Vol. 1, Toulouse, 2006, pp. 265-268.

[16] V. Vapnik. Statistical Learning Theory, Wiley, 1998.

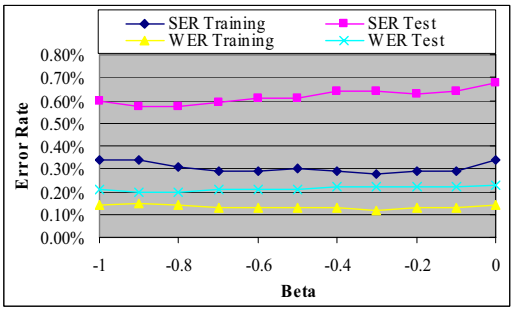

Fig. 1: Recognition error rate as a function of β, which is fixed over MCE training iterations (Method 1).

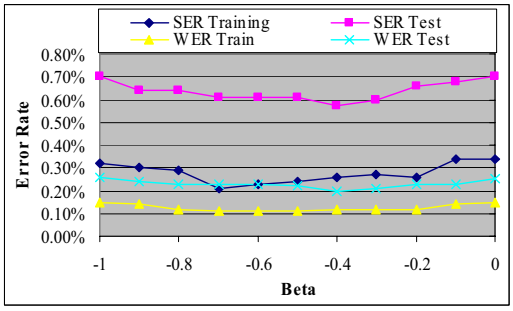

Fig. 2: Recognition error rate as a function of β, which varies over MCE training iterations from 0 to -1 with a decrement of 0.1 (Method 2).

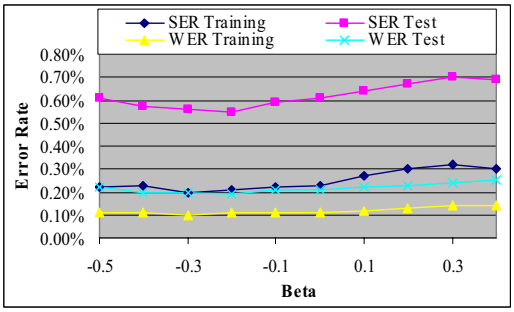

Fig. 3: Recognition error rate as a function of β, which varies over MCE training iterations from 0.4 to -0.5 with a decrement of 0.1 (Method 3).