DISCRIMINATIVE TRAINING OF CDHMMS FOR MAXIMUM RELATIVE SEPARATION MARGIN

Chaojun Liu, Hui Jiang, Xinwei Li

Department of Computer Science and Engineering, York University,
4700 Keele Street, Toronto, Ontario M3J 1P3, CANADA
Email: {cliu, hj, xwli}@cs.yorku.ca

ABSTRACT

In this paper, we propose a new discriminative training method for estimating CDHMM (continuous density hidden Markov model) in speech recognition, based on the principle of maximizing the minimum relative multi-class separation margin. We show that the new training criterion can be formulated as a standard constrained minimax optimization problem. Then we show that the optimization problem can be solved by a GPD (generalized probabilistic descent) algorithm. Experimental results on E-set and Alphabet tasks (ISOLET database) showed that the new training criterion can achieve significant (up to 21%) error rate reduction over the popular MCE (minimum classification error) training method.

1. INTRODUCTION

Discriminative training has been extensively studied over the past decade and been proved quite effective to improve ASR performance over the traditional maximum likelihood (ML) method for HMM-based speech recognition systems. Two most popular discriminative training methods are minimum classification error (MCE) training [5, 2] and maximum mutual information (MMI) training [8, 9]. Despite of their significant progresses, many issues related to discriminative training remain unsolved. One issue reported by many researchers (see [9, 3] and others) is that all discriminative training methods for HMM-based speech recognition suffer the problem of poor generalization capability. In other words, the discriminative training can significantly improve HMMs and leads to a dramatic error reduction on training data but such a significant performance gain can hardly be maintained or generalized in any unseen testing set. Usually only a marginal gain can be achieved over the ML method in a new data set even after discriminative training method is carefully handcrafted for the testing set, especially in large-scale tasks.

To address this problem, we proposed in [4] to estimate HMMs discriminatively based on a new criterion, which is called maximum multi-calss separation margin or large margin estimation (LME). Based on the theoretical results in machine learning, a large margin classifier implies a good generalization power and generally yields much lower generalization errors in new test data as shown in support vector machine and boosting method.

In [1], the authors proposed the so-called Hidden Markov Support Vector machines (HMSVM) for label sequence learning problem. In HMSVM, discrete HMMs (DHMMs) are estimated based on the large margin principle. However, in speech recognition, continuous density HMMs (CDHMM) using Gaussian mixture distributions is the most popular model for speech signals. In this paper, we extend the theoretical study in [4] and propose a modified criterion, namely *large relative margin estimation (LRME)*, to remedy a deficiency in the original criterion in [4]. The remaining of this paper is organized as follows. First, in section 2 we will briefly introduce our original large margin training criterion and then present the modified criterion. Next, in section 3 We will give our solution for estimating large margin CDHMM parameters based on the new criterion using a GPD algorithm [6]. Experimental results will be presented in section 4. Finally a summary will be given in section 5.

2. LARGE RELATIVE MARGIN HMM

In ASR, given any speech utterance X, a speech recognizer will choose the word $\hat{W}$ [1] as output based on the MAP decision rule as follows:

$$\hat{W} = \arg\max_W \ p(W|X) = \arg\max_W \ p(W) \cdot p(X|W) \qquad (1)$$
$$= \arg\max_W p(W) \cdot p(X|\lambda_W) = \arg\max_W \mathcal{F}(X|\lambda_W)$$

where λ_W denotes the HMM representing the word W and $\mathcal{F}(X|\lambda_W)$ is called discriminant function. Here we are only interested in HMM λ_W and assume $p(W)$ is fixed.

For a speech utterance X_i, assuming its true word identity as W_i^T, following [1], the multi-class separation margin for X_i^T is similarly defined as:

$$d(X_i) = \mathcal{F}(X_i|\lambda_{W_i^T}) - \max_{W_j \in \Omega \ W_j \neq W_i^T} \mathcal{F}(X_i|\lambda_{W_j}) \qquad (2)$$

$$= \min_{W_j \in \Omega \ W_j \neq W_i^T} \left[\mathcal{F}(X_i|\lambda_{W_i^T}) - \mathcal{F}(X_i|\lambda_{W_j})\right] \qquad (3)$$

where Ω denotes the set of all possible words. Obviously, if $d(X_i) > 0$, X_i will be correctly recognized; if $d(X_i) \leq 0$, X_i will be incorrectly recognized.

Given a set of training data $\mathcal{D} = \{X_1, X_2, \cdots, X_N\}$, and its transcription $\mathcal{L} = \{W_1^T, W_2^T, \cdots, W_N^T\}$, we can calculate the separation margin (or margin for short) for every utterance in $\mathcal{D}$ based on the definition in eq.(2). Let's define a subset of training utterances, $\mathcal{S}$

$$\mathcal{S} = \{X_i \mid X_i \in \mathcal{D} \text{ and } 0 \leq d(X_i) \leq \gamma\} \qquad (4)$$

where $\gamma > 0$ is a pre-set positive number. $\mathcal{S}$ is called *support vector set* and each utterance in $\mathcal{S}$ is called a support token, which has

[1] Depending on the problem of interest, a word W may be any linguistic unit, e.g., a phoneme, a syllable, a word, a phrase, a sentence, etc..

relatively small positive margin among all utterances in training set $\mathcal{D}$. In other words, all utterances in $\mathcal{S}$ are relatively close to the classification boundaries even though all of them locate in the right decision regions. To achieve a better generalization power, it is desirable to adjust decision boundaries, which are implicitly determined by all models, through optimizing HMM parameters Λ to make all support tokens as far from the decision boundaries as possible, which will result in a robust classifier with better generalization capability. This idea leads to estimating the HMM models Λ based on the criterion of maximizing the minimum margin of all support tokens, which is named as large margin estimation (LME) of HMM.

$$\tilde{\Lambda} = \arg\max_{\Lambda} \min_{X_i \in \mathcal{S}} \quad d(X_i) \tag{5}$$

The HMM models, $\tilde{\Lambda}$, estimated in this way, are called large margin HMMs. Considering eq.(3), large margin HMMs can be equivalently estimated as follows:

$$\tilde{\Lambda} = \arg\max_{\Lambda} \min_{X_i \in \mathcal{S},\, W_j \in \Omega j \neq i} \left[\mathcal{F}(X_i|\lambda_{W_i^T}) - \mathcal{F}(X_i|\lambda_{W_j}) \right] \tag{6}$$

$$= \arg\min_{\Lambda} \max_{X_i \in \mathcal{S},\, W_j \in \Omega j \neq i} \left[\mathcal{F}(X_i|\lambda_{W_j}) - \mathcal{F}(X_i|\lambda_{W_i^T}) \right] \tag{7}$$

subject to contraint $\mathcal{F}(X_i|\lambda_{W_i^T}) - \mathcal{F}(X_i|\lambda_{W_j}) > 0$ for all $X_i \in \mathcal{S}$ and $W_j \in \Omega,\ j \neq i$.

However, such a constraint does not guarantee the existence of a minimax point. As an illustration of this, let's assume a simple case with only two classes $m1$ and $m2$ and there is a support token X close to the decision boundary. If we pull $m1$ and $m2$ together at the same time, we can keep the boundary unchanged but increase the margin defined in eq.(3) as much as we want. As models move toward X, the absolute values of both $\mathcal{F}(X|m1)$ and $\mathcal{F}(X|m2)$ increase, so does the margin as well, although the relative position of X related to the boundary actually doesn't change at all.

There are a few ways to remedy this deficiency in the original LME criterion. One solution is proposed in [7]. In this paper, we propose to change the definition of margin in eq.(2) to be a relative separation margin, defined as:

$$\tilde{d}(X_i) = \min_{W_j \in \Omega\ W_j \neq W_i^T} \left[\frac{\mathcal{F}(X_i|\lambda_{W_i^T}) - \mathcal{F}(X_i|\lambda_{W_j})}{\mathcal{F}(X_i|\lambda_{W_i^T})} \right] \tag{8}$$

If the discriminant functions $\mathcal{F}(\cdot)$ are defined as in eq.(1), for all support tokens in the set $\mathcal{S}$, the dynamic range of the relative margin $\tilde{d}(X_i)$ lies in $[0, 1]$. Since the relative margin is bounded by definition, the maximum value of relative margin always exists. However, in many cases, $\mathcal{F}(X_i|\lambda)$ is defined as the log-likelihood of X_i given model set Λ, so we have $\mathcal{F}(X_i|\lambda_{W_i^T}) < 0$ and $\mathcal{F}(X_i|\lambda_{W_j}) < 0$. To make the relative margin a positive value, we slightly modify its definition as:

$$\tilde{d}(X_i) = \min_{W_j \in \Omega\ W_j \neq W_i^T} \left[\frac{\mathcal{F}(X_i|\lambda_{W_j}) - \mathcal{F}(X_i|\lambda_{W_i^T})}{\mathcal{F}(X_i|\lambda_{W_i^T})} \right] \tag{9}$$

Thus, as $\mathcal{F}(X_i|\lambda_{W_j}) < \mathcal{F}(X_i|\lambda_{W_i^T})$ for correctly recognized tokens in the set $\mathcal{S}$, we always have $d(X_i) > 0$. Similarly we define the support vector set as eq.(4). Therefore, our new training criterion is defined as

$$\hat{\Lambda} = \arg\min_{\Lambda} \max_{X_i \in S,\, W_j \in \Omega, W_j \neq W_i^T} \left[1 - \frac{\mathcal{F}(X_i|\lambda_{W_j})}{\mathcal{F}(X_i|\lambda_{W_i^T})} \right] \tag{10}$$

subject to constraints

$$\mathcal{F}(X_i|\lambda_{W_i^T}) - \mathcal{F}(X_i|\lambda_{W_j}) > 0$$

for all $X_i \in \mathcal{S}$ and $W_j \in \Omega,\ j \neq i$. It is called *large relative margin estimation (LRME)* of HMMs. To solve the above minimax optimization problem, we will derive an iterative approach for CDHMM based on the GPD algorithm.

An intuitive explanation of the large margin estimation (LME) or large relative margin estimation (LRME) can be illustrated by a simple HMM-based classifier for 2-class problem, as shown in Figure 1. By modifying the HMM parameters, we change the classification boundary to make it as far from all training samples as possible. In this way, margin of the classifier will be increased so that its generalization power is improved accordingly.

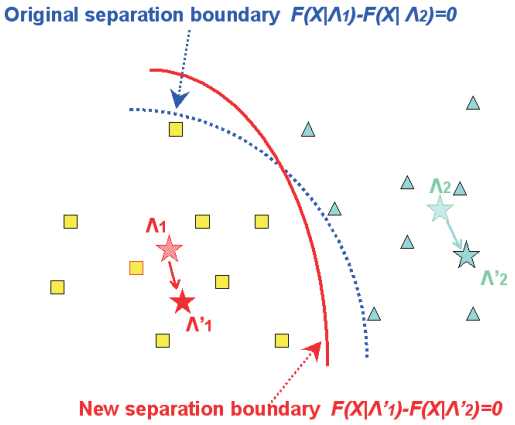

Fig. 1. Illustation of a Large Margin or Large Relative Margin classifier

3. LARGE RELATIVE MARGIN ESTIMATION OF CDHMM USING GPD ALGORITHM

To use GPD algorithm for the large margin optimization in eq.(6), we need to construct a differentiable objective function. We use summation of exponential functions to approximate the maximization in eq.(10) as follows:

$$\max_{X_i \in S,\, W_j \in \Omega,\, W_j \neq W_i^T} \left[1 - \frac{\mathcal{F}(X_i|\lambda_{W_j})}{\mathcal{F}(X_i|\lambda_{W_i^T})} \right]$$

$$\approx \log \left\{ \sum_{X_i \in S,\, W_j \in \Omega,\, W_j \neq W_i^T} \exp \left[\eta d(X_i, \lambda_{W_j}, \lambda_{W_i^T} \right] \right\}^{1/\eta} \tag{11}$$

$$d(X_i, \lambda_{W_j}, \lambda_{W_i^T}) = d_{ij} = 1 - \frac{\mathcal{F}(X_i|\lambda_{W_j})}{\mathcal{F}(X_i|\lambda_{W_i^T})}$$

where $\eta > 1$. As $\eta \to \infty$, the continuous function in the right hand side of eq.(11) will approach the maximization in the left hand side.

Therefore, we define the objective function as:

$$Q(\Lambda) = \frac{1}{\eta} \log \left\{ \sum_{X_i \in S,\, W_j \in \Omega,\, W_j \neq W_i^T} \exp(\eta d_{ij}) \right\} \quad (12)$$

$$= \frac{1}{\eta} \log \left\{ \sum_{X_i \in S} \sum_{W_j \in \Omega, W_j \neq W_i^T} \exp(\eta d_{ij}) \right\} \quad (13)$$

$$= \frac{1}{\eta} \log Q_1 \quad (14)$$

Now, we can use GPD algorithm to adjust Λ to minimize the objective function $Q(\Lambda)$. To maintain HMM model constraints during the optimization process, we need to define the same transformations for model parameters as used in [5]. Then it can be shown that the iterative adjustment of Gaussian means follows

$$\tilde{\mu}_{skl}^m(n+1) = \tilde{\mu}_{skl}^m(n) - \epsilon \frac{\partial Q(\Lambda)}{\partial \tilde{\mu}_{skl}^m}\bigg|_{\Lambda=\Lambda_n} \quad (15)$$

$$\mu_{skl}^m(n+1) = \sigma_{skl}^m \, \tilde{\mu}_{skl}^m(n+1) \quad (16)$$

where $\mu_{skl}^m(n+1)$ is the l-th dimension of Gaussian mean vector for the k-th mixture component of state s of HMM model m at $n+1$ iteration.

$$\frac{\partial Q(\Lambda)}{\partial Q_1} = \frac{1}{\eta} \frac{1}{Q_1} \quad (17)$$

$$\frac{\partial Q_1}{\partial \tilde{\mu}_{skl}^m} = \sum_{X_i \in S} \left\{ \sum_{W_j \in \Omega, W_j \neq W_i^T} \eta \, \exp(\eta d_{ij}) \frac{\partial d_{ij}}{\partial \tilde{\mu}_{skl}^m} \right\}$$

$$= \sum_{X_i \in S} \left\{ \Big[\delta(W_i^T - m) \frac{1}{\mathcal{F}^2(X_i|\lambda_m)} \frac{\partial \mathcal{F}(X_i|\lambda_m)}{\partial \tilde{\mu}_{skl}^m} \sum_{W_j \in \Omega, W_j \neq m} \eta \, \mathcal{F}(X_i|\lambda_{W_j}) \exp(\eta d_{ij}) \Big] - (1 - \delta(W_i^T - m)) \frac{1}{\mathcal{F}(X_i|\lambda_{W_i^T})} \eta \, \exp(\eta d_{ij}) \frac{\partial \mathcal{F}(X_i|\lambda_m)}{\partial \tilde{\mu}_{skl}^m} \right\} \quad (18)$$

where M is the total number of hmm models in Λ. $W_i^T = m$ if the true model for utterance X_i is the m-th model in the model set Λ. As

$$\mathcal{F}(X_i|\lambda_m) = \log \, L(X_i, \lambda_m) \approx \log \, L(X_i, q; \lambda_m)$$

$$= \sum_{t=1}^{T} \Big[\log \, a_{q_{t-1} q_t}^m + \log \, b_{q_t}^m(x_t) \Big] + \log \, \pi_{q_0}^m \quad (19)$$

$$b_j^m(x_t) = \sum_{k=1}^{K} c_{jk}^m \mathcal{N}\big[x_t; \mu_{jk}^m, R_{jk}^m\big] \quad (20)$$

so,

$$\frac{\partial \mathcal{F}(X_i|\lambda_m)}{\partial \tilde{\mu}_{skl}} = \sum_{t=1}^{T} \delta(q_t - s) \frac{\partial \log \, b_s^m(x_t)}{\partial \tilde{\mu}_{skl}} \quad (21)$$

where

$$\frac{\partial \log \, b_s^m(x_t)}{\partial \tilde{\mu}_{skl}^m} = c_{sk}^m (2\pi)^{-\frac{D}{2}} \big\| R_{sk}^m \big\|^{-\frac{1}{2}} (b_s^m(x_t))^{-1}$$

$$\left(\frac{x_{tl} - \mu_{skl}^m}{\sigma_{skl}^m} \right) \exp \left\{ -\frac{1}{2} \sum_{l=1}^{D} \left(\frac{x_{tl} - \mu_{skl}^m}{\sigma_{skl}} \right)^2 \right\} \quad (22)$$

D is the dimension of feature vectors. R_{sk}^m is the covariance matrix for state s and Gaussian mixture component k for hmm model m. Here we assume it is diagonal. q is the best state sequence obtained by aligning X_i using hmm model λ_m.

Combining equations from (17) to (22), we can easily obtain $\partial Q(\Lambda)/\partial \tilde{\mu}_{skl}^m$ for eq.(15). Similar derivations for the variances, mixture weights and transition probabilities can be easily accomplished. However, we only updated mean vectors in the work reported in this paper.

4. EXPERIMENTAL RESULTS

The LRME algorithm was tested on two isolated-word tasks. The first one is E-set (B, C, D, E, G, P, T, V, Z), the second one is alphabet (letters A-Z) set. The OGI-ISOLET database was used. The ML (maximum likelihood) baseline was built using HTK toolkit. The ISOLET set 1-4 first production of each speaker was used as the training set. There are 1080 utterances for E-set and 3120 utterances for alphabet set, respectively. Both productions of every speaker in ISOLET set 5 were used as the testing set. There are 540 utterances for E-set and 1560 for alphabet set in the testing data. The data sampling rate is 16K Hz. Acoustic feature vectors are of standard 39 dimensions including 12 MFCC, the normalized energy, and their first and second order time derivatives. Each letter was represented by a whole-word HMM model with 16 states. Different number of Gaussian mixture components are experimented. The MCE model training uses the best ML model as the seed model. The LRME model training uses one of the best MCE models as the seed model. The tables 1 and 2 give a performance comparison of the best results obtained by different training criteria.

Table 1. Results (word accuracy %) of different training criteria on E-set test data.

	1-mixture	2-mixture	4-mixture
ML	85.56	90.56	91.48
MCE	91.48	94.07	**94.44**
LRME	93.52	95.00	**95.19**

Table 2. Results (word accuracy %) of different training criteria on Alphabet test data.

	1-mixture	2-mixture	4-mixture
ML	93.14	94.94	95.38
MCE	95.58	95.96	**96.09**
LRME	95.64	96.60	**96.92**

It is clearly demonstrated that LRME achieved the best results on both tasks. On E-set, LRME-trained model obtained the word accuracy of 95.19%, which is 13% less error than the best MCE-trained model with an accuracy of 94.44%. On Alphabet set, LRME achieved a 96.92% word accuracy, which is 21% less error than the best MCE model with an accuracy of 96.09%.

Figure 2 plots the recognition accuracy, approximate margin , i.e., $-Q(\Lambda)$, where$Q(\Lambda)$ is given in eq.(12)), and true margin on the test data, i.e., $d(X_i)$ in eq.(9), as a function of the number

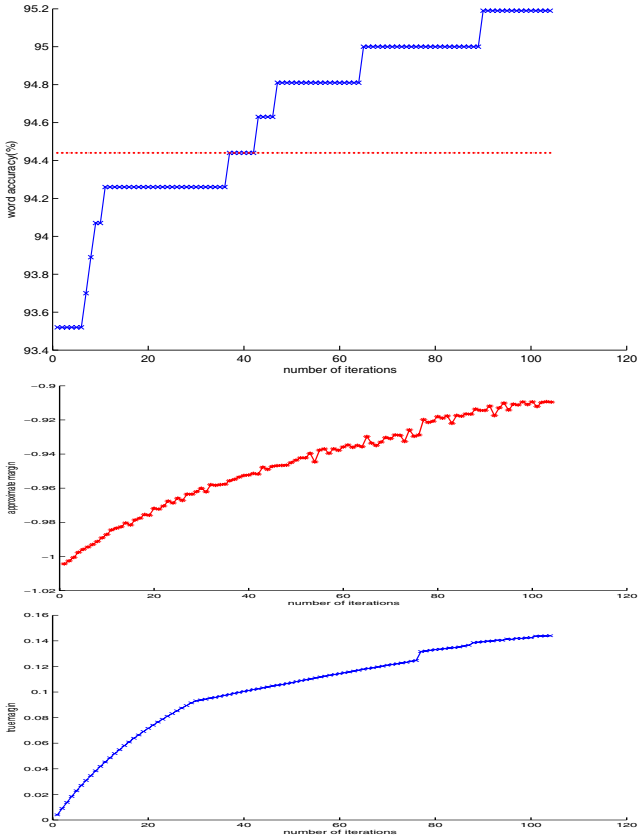

Fig. 2. Curves for LRME training of a 4-mixture model on E-set task. Top figure is word accuracy of LRME model on testing set (flat line is the best MCE accuracy level for comparison). The middle figure is the curve of approximate margin which was being maximized during LRME-training. The bottom one is the curve of corresponding true margins.

of iterations of the LRME training procedure. As the seed model (obtained from previous MCE training) already reaches 100% accuracy on the training set and the LRME training process keeps the accuracy unchanged, so it was not plot in the figure. We can see from the curves that with the number of iterations going up, the approximate margin keeps increasing, which is consistent with the goal of GPD optimization. Meanwhile the recognition accuracy on the testing set keeps increasing (or unchanged for a short period). After 40 iterations, the LRME model reaches the accuracy level of the best MCE model on the testing set. After 90 iterations, the LRME training achieves 95.19% accuracy on the testing set, representing a 13% reduction in recognition error. Also we can see that the true margin keeps increasing accordingly and it is greater than the approximate margin. It can be proved that the approximate margin is a lower bound of the true margin. Our study shows that the larger η, the approximate margin get closer to the true margin. But a too large η may make the LRME estimation very sensitive to an outlier training sample. In our experiments, we choose η as 10.

Similar to MCE estimation, there are a few parameters (ϵ in eq.(15), η in eq.(12), γ in eq.(4)) that affect the convergence of LRME/GPD estimation. A detailed study of their effects will be given in another paper due to lack of space.

Another issue is that the current LRME training uses only support tokens, which are a subset of correctly recognized tokens. In case there is any recognition error in the training set, a different algorithm, which is similar to MCE formulation, was proposed in [4] to handle the set of error tokens. However, in the work reported here, there is no recognition error in the training set, so we are not concerned about it.

5. SUMMARY

In this paper, we have proposed a new training method, large relative margin estimation(LRME), for continuous density HMM based speech recognition. The LRME approach aims at improving the poor generalization capability of existing discriminative training methods. Motivated by large margin classifier in machine learning, the new training criterion is trying to maximize the minimum multi-class relative separation margin. The formulation of GPD-based LRME is given. We investigated its performance on two speaker-independent isolated-word tasks. The LRME method provides up to 21% reduction in error rate, compared to the MCE method. Further research and experiments on continuous speech and sub-word based system are in progress.

6. REFERENCES

[1] Y. Altun and T. Hfomann, "Large margin methods for label sequence learning," *Eurospeech 2003*, pp.993–996, Geneva, Switzerland, Sep. 2003.

[2] H. Jiang, O. Siohan, F. Soong and C.-H. Lee, "A dynamic in-search discriminative training approach for large vocabulary speech recognition," *ICASSP'2002*, pp.I-113-116, Orlando, Florida, May 2002.

[3] H. Jiang, F. Soong and C.-H. Lee, "A dynamic in-search data selection method with its applications to acoustic modeling and utterance verification," *to appear in IEEE Trans. on SAP*, June 2003.

[4] H. Jiang, "Discriminative Training for Large Margin HMMs", *Technical Report CS-2004-01, CSE Department, York University*, March 2004.
(http://www.cs.yorku.ca/techreports/2004/CS-2004-01.html)

[5] B.-H. Juang, W. Chou and C.-H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition," *IEEE Trans. on Speech and Audio Processing (SAP)*, pp.257-265, Vol.5, No.3, May 1997.

[6] S. Katagiri, B.-H. Juang and C.-H. Lee, "Pattern recognition using a generalized probabilistic descent method," *Proceedings of the IEEE*, Vol. 86, No. 11, pp.2345-2373, Nov. 1998.

[7] X. Li, H. Jiang and C. Liu, "Large Margin HMM for Speech Recognition" *submitted to ICASSP'2005*, March, 2005.

[8] Y. Normandin, R. Cardin and R. Demori, "High-performance connected digit recognition using maximum mutual information estimation," *IEEE Trans. on SAP*, Vol. 2, No. 2, Apr. 1994.

[9] P.C. Woodland and D. Povey, "Large Scale Discriminative Training of hidden Markov models for speech recognition," *Computer Speech & Language*, pp.25-47, Vol. 16, No. 1, January 2002.