YORK
UNIVERSITÉ
UNIVERSITY

Discriminative Training for Large Margin HMMs

Hui Jiang

Technical Report CS-2004-01

March 30, 2004

Department of Computer Science

4700 Keele Street North York, Ontario M3J 1P3 Canada

Discriminative Training for Large Margin HMMs

Hui Jiang

Department of Computer Science and Engineering, York University,
4700 Keele Street, Toronto, Ontario M3J 1P3, CANADA

Correspondence:

Prof. Hui Jiang, Department of Computer Science and Engineering, York University,
4700 Keele Street, Toronto, Ontario M3J 1P3, CANADA

Phone: (416)736-2100 x33346 (Hui Jiang)

Fax: (416)736-5872 (Hui Jiang)

Email: hj@cs.yorku.ca

Abstract

In this report, motivated by large margin classifiers in machine learning, we propose a new discriminative training criterion for estimating CDHMM (continuous density hidden Markov model) in speech recognition based on the principle of maximizing the minimum multi-class separation margin. In this report, we first show that this maximum margin model estimation problem can be formulated as a standard constrained minimax optimization problem. Alternatively, we also show that the estimation problem can be solved by a GPD (generalized probabilistic descent) algorithm if we approximate the objective function by a continuous and differentiable function, such as summation of exponential functions. In this report, we also propose a method to handle classification errors in training set in maximum margin estimation by using them to optimize a separate objective function which is similar to that in the MCE (minimum classification error) formulation.

1 Introduction

The most successful modeling approach to automatic speech recognition (ASR) is to use a set of HMMs as the acoustic models of subword or whole-word speech units and to use the statistical N-gram model as language models for words and/or word classes. All the model parameters, including HMMs and N-gram models, are estimated from a large amount of training data. As for HMM-based acoustic models, the dominant estimation method is the Baum-Welch algorithm based on the maximum likelihood (ML) criterion. The ML estimation methods of HMM parameters have been developed for a variety of HMM types in the last two decades, e.g., in [6, 19, 14] and many others. As an alternative to the standard Maximum Likelihood (ML) estimation, discriminative training has also been extensively studied for HMM-based automatic speech recognition (ASR). Some discriminative training methods aim to improve model separation among all models, such as maximum mutual information (MMI) training [4], conditional maximum likelihood estimation (CMLE) [21] and H-criteria[8]. Other methods try to directly reduce the recognition error rate on training data, such as corrective training [5], minimum empirical error rate training [20] and MCE (minimum classification error) training [15, 16, 17]. Among these approaches, the MCE formulation has been regarded as one of the most successful methods. In the MCE, the empirical error rate on training data is approximated by a smoothed and differentiable objective function and then the GPD (generalized probabilistic descend) algorithm [17] is used to minimize the objective function with respect to all HMM parameters.

Discriminative training has been found quite effective to improve ASR performance over the ML method in small or medium vocabulary ASR tasks (see [22, 16]). However, no significant gain in performance had been demonstrated in any large-scale ASR tasks until very recently. In [25], the MMI method was applied to the switchboard task and some moderate but consistent improvements over the conventional ML method were observed in their experiments while in [11, 12] the MCE method was extended to a large-scale continuous speech recognition task, e.g., the DARPA communicator task, and similarly a slight gain was achieved over the best ML-trained HMMs. Despite of these significant progresses, many issues related to discriminative training still remain unsolved. For instance, as reported by many researchers (see [25, 12] and others), 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 new unseen test 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 test set, especially in large-scale tasks. So far, the two major discriminative training criteria, namely maximum mutual information (MMI) and minimum classification error (MCE), have been extensively studied for speech recognition. Intuitively, a better discriminative training criterion will lead to a better generalization capability. Motivated by some recent advances in machine learning about large margin classifiers, in this report, I propose to estimate HMMs discriminatively based on a new criterion, such as maximum separation margin, as in other large margin classifiers. 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 (SVM) and boosting method.

In [1, 2], 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. As shown in [1, 2], estimation of DHMMs for large margin turns out to be a quadratic programming problem under some constraints. The problem can be solved by many standard optimization software tools similarly as the standard support vector machine (SVM). However, in automatic speech recognition (ASR), Gaussian mixture continuous density HMM (CDHMM) is the most popular model for modeling speech signals. In this paper, we study how to estimate

CDHMM based on the above large margin principle for speech recognition. I will show that the proposed approach is similar to the standard GPD-based MCE discriminative training which has been extensively studied for years but it is based on a new discrimination criterion, namely maximizing the minimum margin of HMMs in classification.

2 Large Margin HMMs for ASR

In ASR, given any speech utterance X, a speech recognizer will choose the word $\hat{W}$[1] as output based on the plug-in MAP decision rule [10, 18] as follows:

$$\begin{aligned} \hat{W} &= \arg\max_{W} \; p(W|X) = \arg\max_{W} \; p(W) \cdot p(X|W) \\ &= \arg\max_{W} p(W) \cdot p(X|\lambda_W) = \arg\max_{W} \mathcal{F}(X|\lambda_W) \end{aligned} \tag{1}$$

where λ_W denotes the HMM representing the word W and $\mathcal{F}(X|\lambda_W) = p(W) \cdot p(X|\lambda_W)$ is called discriminant function. In this work, we are only interested in how to estimate HMM λ_W and assume language model used to calculate $p(W)$ is fixed.

For a speech utterance X_i, assuming its true word identity as W_i, following [1, 2], the multi-class separation margin for X_i is similarly defined as:

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

where Ω denotes the set of all possible words. Apparently, the above equation (2) can be re-arranged into:

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

Obviously, if $d(X_i) \leq 0$, X_i will be incorrectly recognized by the current HMM set, denoted as Λ; if $d(X_i) > 0$, X_i will be correctly recognized by the models Λ.

Given a set of training data $\mathcal{D} = \{X_1, X_2, \cdots, X_T\}$, we usually know the true word identities for all utterances in $\mathcal{D}$, denoted as $\mathcal{L} = \{W_1, W_2, \cdots, W_T\}$. Thus, we can

[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..

calculate the separation margin (or margin for short hereafter) for every utterance in $\mathcal{D}$ based on the definition in eq.(2) or (3). If we want to estimate the HMM parameters Λ, one desirable estimation criterion is to minimize the total number of utterances in whole training set which have negative margin as in the standard MCE estimation [15, 16]. Furthermore, motivated by the large margin principle in machine learning, even for those utterances which all have positive margin, we may still want to maximize the minimum margin among them towards a HMM-based large margin classifier for ASR. Based on the machine learning theory, a large margin classifier usually leads to much lower generalization error rate in a new test set and shows a more robust and better generalization capability. In this work, we will study how to estimate HMMs for speech recognition based on the above-mentioned principle of maximizing minimum margin.

First of all, from all utterances in $\mathcal{D}$, we need to identify a subset of utterances $\mathcal{S}$ as:

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

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

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

where the above maximization and minimization are performed subject to the constraints that $d(X_i) > 0$ for all $X_i \in \mathcal{S}$. 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:

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

subject to

$$\mathcal{F}(X_i|\lambda_{W_i}) - \mathcal{F}(X_i|\lambda_{W_j}) > 0 \tag{7}$$

for all $X_i \in \mathcal{S}$ and $W_j \in \Omega$ and $j \neq i$.

Finally, the above optimization can be converted into a standard minimax optimization problem as:

$$\tilde{\Lambda} = \arg\min_{\Lambda} \max_{X_i \in \mathcal{S}} \max_{W_j \in \Omega \; j \neq i} \left[\mathcal{F}(X_i|\lambda_{W_j}) - \mathcal{F}(X_i|\lambda_{W_i}) \right] \tag{8}$$

where the minimax optimization is subject to the following constraints:

$$\mathcal{F}(X_i|\lambda_{W_j}) - \mathcal{F}(X_i|\lambda_{W_i}) < 0 \tag{9}$$

for all $X_i \in \mathcal{S}$ and $W_j \in \Omega$ and $W_j \neq W_i$.

Obviously, the above minimax optimization can be numerically solved by many optimization software tools. In this report, instead we will approximate the objective function in the above minimax optimization and then derive an iterative optimization approach for CDHMM based on the GPD algorithm[17].

3 Large Margin Estimation of CDHMM

In this section, let's describe how to solve the above optimization problem in eq.(8) for CDHMM in speech recognition based on a GPD iterative optimization algorithm. At first, we assume each speech unit, e.g., a word W, is modeled by an N-state CDHMM with parameter vector $\lambda = (\pi, A, \theta)$, where π is the initial state distribution, $A = \{a_{ij} | 1 \leq i, j \leq N\}$ is transition matrix, and θ is parameter vector composed of mixture parameters $\theta_i = \{\omega_{ik}, m_{ik}, r_{ik}\}_{k=1,2,\cdots,K}$ for each state i, where K denotes number of Gaussian mixtures in each state. The state observation p.d.f. is assumed to be a mixture of multivariate Gaussian distribution:

$$\begin{aligned} p(\mathbf{x}|\theta_i) &= \sum_{k=1}^{K} \omega_{ik} \cdot \mathcal{N}(\mathbf{x}|m_{ik}, r_{ik}) \\ &= \sum_{k=1}^{K} \omega_{ik} \cdot (2\pi)^{-D/2} |r_{ik}|^{1/2} \ \exp\left[-\frac{1}{2}(x - m_{ik})^t r_{ik} (x - m_{ik}) \right] \end{aligned} \tag{10}$$

where mixture weights ω_{ik}'s satisfy the constraint $\sum_{k=1}^{K} \omega_{ik} = 1$. In many cases, we prefer to use multivariate Gaussian distribution with diagonal precision matrix. Thus, the above state observation p.d.f. is simplified as:

$$p(\mathbf{x}|\theta_i) = \sum_{k=1}^{K} \omega_{ik} \mathcal{N}(\mathbf{x}|m_{ik}, r_{ik}) = \sum_{k=1}^{K} \omega_{ik} \prod_{d=1}^{D} \sqrt{\frac{r_{ikd}}{2\pi}} e^{-\frac{1}{2} r_{ikd} (x_d - m_{ikd})^2} \tag{11}$$

Given any speech utterance $X_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \cdots, \mathbf{x}_{iR}\}$, let $\mathbf{s} = \{s_1, s_2, \cdots, s_R\}$ be the unobserved state sequence, and $\mathbf{l} = \{l_1, l_2, \cdots, l_R\}$ be the associated sequence of the unobserved mixture component labels, the discriminant function based on the word model λ_{W_j}, $\mathcal{F}(X_i|\lambda_{W_j})$, can be expressed as:

$$\mathcal{F}(X_i|\lambda_{W_j}) = \sum_{\mathbf{s}} \sum_{\mathbf{l}} \left\{ \pi_{s_1} \omega_{s_1 l_1} \mathcal{N}(\mathbf{x}_{i1}|m_{s_1 l_1}, r_{s_1 l_1}) \prod_{t=2}^{R} a_{s_{t-1} s_t} \cdot \omega_{s_t l_t} \cdot \mathcal{N}(\mathbf{x}_{it}|m_{s_t l_t}, r_{s_t l_t}) \right\} \cdot p(W_j) \tag{12}$$

where the summations are taken over all possible state and mixture component label sequences. But, if we can use the Viterbi method to approximate the above summation with the single optimal Viterbi path and assume the optimal Viterbi path is denoted as $\mathbf{s}^* = \{s_1^*, s_2^*, \cdots, s_R^*\}$ and $\mathbf{l}^* = \{l_1^*, l_2^*, \cdots, l_R^*\}$, then we have

$$\mathcal{F}(X_i|\lambda_{W_j}) \approx \pi_{s_1^*} \omega_{s_1^* l_1^*} \mathcal{N}(\mathbf{x}_{i1}|m_{s_1^* l_1^*}, r_{s_1^* l_1^*}) \prod_{t=2}^{R} a_{s_{t-1}^* s_t^*} \cdot \omega_{s_t^* l_t^*} \cdot \mathcal{N}(\mathbf{x}_{it}|m_{s_t^* l_t^*}, r_{s_t^* l_t^*}) \cdot p(W_j) \tag{13}$$

In most cases, it is more convenient to represent the discriminant function $\mathcal{F}(X_i|\lambda_{W_j})$ in the logarithm scale. Assume we adopt diagonal precision matrices for all Gaussian mixtures, we have

$$\begin{aligned}\mathcal{F}(X_i|\lambda_{W_j}) &\approx \log p(W_j) + \log \pi_{s_1^*} + \sum_{t=2}^{R} \log a_{s_{t-1}^* s_t^*} + \prod_{t=1}^{R} \log \omega_{s_1^* l_1^*} \\ &+ \frac{1}{2} \sum_{t=1}^{R} \sum_{d=1}^{D} \left[\log r_{s_t^* l_t^* d} - r_{s_t^* l_t^* d} (\mathbf{x}_{itd} - m_{s_t^* l_t^* d})^2 \right] \end{aligned} \quad (14)$$

To construct a differentiable objective function for the large margin optimization in eq.(8), we need to approximate the *max* operation with a continuous and differentiable function. Usually either exponential or power function can be used for this purpose. Here assume we use summation of exponential functions to approximate the maximization in eq(8) as follows:

$$\begin{aligned} &\max_{X_i \in \mathcal{S} \;\, W_j \in \Omega \; j \neq i} \left[\mathcal{F}(X_i|\lambda_{W_j}) - \mathcal{F}(X_i|\lambda_{W_i}) \right] \\ \approx \;& \log \left[\sum_{X_i \in \mathcal{S} \; W_j \in \Omega \; j \neq i} \exp \left[\eta \cdot \mathcal{F}(X_i|\lambda_{W_j}) - \eta \cdot \mathcal{F}(X_i|\lambda_{W_i}) \right] \right]^{1/\eta} \end{aligned} \quad (15)$$

where $\eta > 1$. As $\eta \to \infty$, the continuous function in the right hand side of eq.(15) will approach the maximization in the left hand side. In practice, we can choose η as a constant larger than 1.

Therefore, we derive the objective function for large margin estimation (LME) of CDHMM as following:

$$Q_1(\Lambda) = -\frac{1}{\eta} \log \left[\sum_{X_i \in \mathcal{S} \; W_j \in \Omega \; j \neq i} \exp \left[\eta \cdot \mathcal{F}(X_i|\lambda_{W_j}) - \eta \cdot \mathcal{F}(X_i|\lambda_{W_i}) \right] \right] \quad (16)$$

where $\eta > 1$ is set as a constant beforehand, and all $\mathcal{F}(X_i|\lambda_{W_j})$ are given by eq.(12) or (14). If we adopt the Viterbi approximation in eq.(12), the above $Q_1(\cdot)$ has a relatively simple form. Usually, a gradient descent iteration method must be used to minimize $Q_1(\Lambda)$ with respect to all CDHMM parameters Λ in order to derive the large margin estimation of CDHMM as in eq.(8). The minimization is subject to all constraints given in eq.(9). In practice, all these inequality constraints can be cast as interior penalty functions as in [3].

Here let's first consider a simple case, where we only re-estimate mean vectors of CDHMMs based on the large margin principle while keeping all other CDHMM parameters constant during the large margin estimation. For any utterance X_i in the support token set $\mathcal{S}$, if we assume its true model is λ_i, then we check for other models λ_j $(j \neq i)$ to include all pairs of the utterance X_i (assume its true model is λ_i) and a hypothesized incorrect model λ_j for large margin model estimation as long as they meet the condition $0 < \mathcal{F}(X_i|\lambda_i) - \mathcal{F}(X_i|\lambda_j) \leq \epsilon$, where ϵ is a pre-set threshold. For simplicity, we use the Viterbi approximation in evaluating both $\mathcal{F}(X_i|\lambda_i)$ and $\mathcal{F}(X_i|\lambda_j)$. For $\mathcal{F}(X_i|\lambda_i)$, let's assume the optimal Viterbi path is $\mathbf{s}' = \{s'_1, s'_2, \cdots, s'_T\}$ and $\mathbf{l}' = \{l'_1, l'_2, \cdots, l'_T\}$. Similarly, we assume the optimal path is $\mathbf{s}'' = \{s''_1, s''_2, \cdots, s''_T\}$ and $\mathbf{l}'' = \{l''_1, l''_2, \cdots, l''_T\}$ when evaluating $\mathcal{F}(X_i|\lambda_j)$. Since we are only considering to estimate mean vectors of CDHMMs, we can re-write $\mathcal{F}(X_i|\lambda_i)$ and $\mathcal{F}(X_i|\lambda_j)$ according to eq.(14) as follows:

$$\mathcal{F}(X_i|\lambda_i) \approx C' - \frac{1}{2}\sum_{t=1}^{T}\sum_{d=1}^{D} r_{s'_t l'_t d}(x_{itd} - m_{s'_t l'_t d})^2 \tag{17}$$

$$\mathcal{F}(X_i|\lambda_j) \approx C'' - \frac{1}{2}\sum_{t=1}^{T}\sum_{d=1}^{D} r_{s''_t l''_t d}(x_{itd} - m_{s''_t l''_t d})^2 \tag{18}$$

where C' and C'' are two constants independent from mean vectors. In this case, the discriminant functions $\mathcal{F}(X_i|\lambda_i)$ and $\mathcal{F}(X_i|\lambda_j)$ can be represented as a summation of some quadratic functions related to mean values of CDHMMs. Then we can represent the decision margin $\mathcal{F}(X_i|\lambda_i) - \mathcal{F}(X_i|\lambda_j)$ as:

$$\mathcal{F}(X_i|\lambda_i) - \mathcal{F}(X_i|\lambda_j) \approx C - \frac{1}{2}\sum_{t=1}^{T}\sum_{d=1}^{D}\left[r_{s'_t l'_t d}(x_{itd} - m_{s'_t l'_t d})^2 - r_{s''_t l''_t d}(x_{itd} - m_{s''_t l''_t d})^2\right] \tag{19}$$

where $C = C' - C''$.

Obviously, we can substitute the above decision margins in eq.(19) for all support token in the set $\mathcal{S}$ (as defined by $0 < \mathcal{F}(X_i|\lambda_i) - \mathcal{F}(X_i|\lambda_j) \leq \epsilon$) into the objective function in eq.(16) to optimize all HMM parameters. Alternatively, the functions as shown in eq.(19) can be viewed as component functions as well as constraints in a minimax optimization problem when using a standard optimization software tool to solve the minimax optimization problem in eq.(8).

4 Handling Recognition Errors

In case there is any recognition error in training set $\mathcal{D}$, e.g., if an utterance $X_i (X_i \in \mathcal{D})$ is mis-recognized by the current CDHMM set, Λ, then the margin of this utterance, $d(X_i)$, is negative. Obviously, the above large margin estimation can only be done for those utterances with positive margins. It does not make sense to optimize models as above for any utterance with negative margin value. If $d(X_i) < 0$, X_i can not be used to maximize the objective function $Q_1(\Lambda)$ to optimize the HMM set Λ as in the above. Here, we propose to handle these mis-recognized utterances separately from the support tokens in the set $\mathcal{S}$. Based on the current CDHMM set Λ, we first identify all mis-recognized utterances, which all have negative margins, as the *error* set $\mathcal{E}$:

$$\mathcal{E} = \{X_i \mid X_i \in \mathcal{D} \text{ and } d(X_i) \leq 0\} \tag{20}$$

For utterances in $\mathcal{E}$, following the MCE training [15, 17], we optimize CDHMM parameters, Λ, to minimize the total number of utterances in $\mathcal{E}$. However, in practice, the total number count of utterances in $\mathcal{E}$ must be smoothed by plugging the margin into the following sigmoid function:

$$l(d(X_i)) = \frac{1}{1 + \exp[\gamma \cdot d(X_i)]} \tag{21}$$

where $\gamma > 1$ is a constant to control the slope of the sigmoid function. As in the MCE formulation[15, 16, 17], the *max* in the definition of margin $d(X_i)$ in eq.(2) need to be approximated by summation of exponential functions. Finally, the smoothed count of total mis-recognized utterances in $\mathcal{E}$ can be expressed as:

$$\begin{aligned} Q_2(\Lambda) &= \sum_{X_i \in \mathcal{E}} l(d(X_i)) = \sum_{X_i \in \mathcal{E}} \frac{1}{1+\exp[\gamma \cdot d(X_i)]} \\ =&\sum_{X_i \in \mathcal{E}} \frac{1}{1+\exp\left[\gamma \cdot \left[\mathcal{F}(X_i|\lambda_{W_i}) - \frac{1}{\eta_2}\log\left(\sum_{W_j \in \Omega\ j\neq i} \exp\left[\eta_2 \cdot \mathcal{F}(X_i|\lambda_{W_j})\right]\right)\right]\right]} \end{aligned} \tag{22}$$

The GPD algorithm can be used to minimize the above objective function with respect to all model parameters, Λ.

5 A Training Algorithm for Large Margin CDHMM

Given a training set $\mathcal{D}$, we can estimate the whole CDHMM set, Λ, to minimize the following new objective function,$Q(\Lambda)$, which is a weighed linear combination of $Q_1(\Lambda)$ and $Q_2(\Lambda)$. Thus,

$$Q(\Lambda) = \kappa \cdot Q_1(\Lambda) + Q_2(\Lambda) \tag{23}$$

where $\kappa > 0$ is a parameter to make a good balance between $Q_1(\Lambda)$ and $Q_2(\Lambda)$. The optimal value for κ can be selected experimentally. Apparently, the objective function, $Q(\Lambda)$, can be optimized by using any gradient descent algorithm, such as generalized probabilistic descent (GPD) algorithm in [15], with respect to CDHMM parameters Λ.

Alternatively, we can also optimize the objective function $Q_1(\Lambda)$ over the set $\mathcal{S}$ and $Q_2(\Lambda)$ over the set $\mathcal{E}$ separately and iteratively. One algorithm to estimate large margin CDHMM in such a way is shown as Algorithm 1.

References

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

Algorithm 1 GPD-based Estimation Algorithm for Large Margin CDHMM

```
Model estimation based on maximum likelihood criterion $\Rightarrow \Lambda^{(0)}$;
set $n = 0$;
repeat
   Identify the error set $\mathcal{E}$ based on the current model $\Lambda^{(n)}$;
   Use GPD to update model to minimize the objective function $Q_2 \Rightarrow \Lambda^{(n+1)}$;
   n=n+1;
   Identify the support set $\mathcal{S}$ based on the current model $\Lambda^{(n)}$;
   Use GPD to update model to minimize the objective function $Q_1 \Rightarrow \Lambda^{(n+1)}$;
   n=n+1;
until Some converge conditions are met.
```

[2] Y. Altun, I. Tsochantaridis and T. Hofmann, " Hidden Markov Support Vector Machines," *Proc. of the 20th International Conference on Machine Learning (ICML-2003)*, Washington DC, 2003.

[3] M. Avriel, *Nonlinear programming: analysis and methods*, Prentice-Hall, Inc., 1976.

[4] L.R. Bahl, P.F. Brown, P.V. De Souza and R.L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," *Proc. of ICASSP-86*, pp.49-52, Tokyo, Japan, 1986.

[5] L. R. Baul, P.F. Brown, P.V. De Souza and R.L. Mercer, "Estimating Hidden Markov model parameters so as to maximize speech recognition accuracy," *IEEE Trans. on Speech and Audio Processing*, Vol. 1, No. 1, pp.77-83, 1983.

[6] L. E. Baum, T. Petrie, G. Soules and N. Weiss, "A maximimization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, Vol. 41, pp. 164-171, 1970.

[7] , C. J. C. Burges, "A Tutorial on Support Vector Machine for Pattern Recognition," *Data Mining and Knowledge Discovery*, No. 2, pp.121-167, 1998.

[8] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, D. Nahamoo and M. A. Picheny, "Decoder selection based on cross-entropies," *Proc. of ICASSP-88* New York, pp.20-23, 1988.

[9] B. S. Gottfried, *Introduction to optimization theory*, Prentice-Hall, Inc., 1973.

[10] H. Jiang, K. Hirose and Q. Huo, "Robust speech recognition based on Bayesian prediction approach", *IEEE Trans. on Speech and Audio Processing*, pp. 426-440, Vol. 7, No.4, July 1999

[11] H. Jiang, O. Siohan, F. Soong and C.-H. Lee, "A dynamic in-search discriminative training approach for large vocabulary speech recognition," Proc. of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2002), pp.I-113-116, Orlando, Florida, May 2002.

[12] 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 Speech and Audio Processing*, June 2003.

[13] H. Jiang, "Confidence Measures for Speech Recognition: A Survey", *Technical Report CS-2003-06*, Department of Computer Science, York University, June 2003. (submitted to *IEEE Signal Processing Magazine*)

[14] B.-H. Juang, S. E. Levinson and M. M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. on Information Theory*, Vol. IT-32, No. 2, pp.307-309, 1986.

[15] B.-H. Juang and S. Katagiri, “Discriminative Learning for Minimum Error Training,” *IEEE Trans. on Acoustic, Speech, Signal Processing*, Vol. 40, pp.3043-3054, No. 12, Dec. 1992.

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

[17] 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.

[18] C.-H. Lee and Q. Huo, ”On adaptive decision rules and decision parameter adaptation for automatic speech recognition,” *Proc. of the IEEE*, Vol.88, No. 8, pp.1241-1296, Aug. 2000.

[19] L. R. Liporace, ”Maximum likelihood estimation for multivariate observations of Markov sources,” *IEEE Trans. on Information Theory*, Vol. IT-28, No. 5, pp.729-734, 1982.

[20] A. Ljolje, Y. Ephraim and L. R. Rabiner, ” Estimation of hidden Markov model parameters by minimizing empirical error rate,” *Proc. of ICASSP-90*, pp.709-712, 1990.

[21] A. Nadas, D. Nahamoo and M. A. Picheny, ”On a model-robust training method for speech recognition,” *IEEE Trans. on Acoustic, Speech and Signal Processing*, Vol. 36, pp.1432-1436, Sep. 1988.

[22] Y. Normandin, R. Cardin and R. Demori, ”High-performance connected digit recognition using maximum mutual information estimation,” *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, Apr. 1994.

[23] A. J. Smola, P. Bartlett, B. Scholkopf and D. Schuurmans (ed.), *Advances in Large Margin Classifiers*, the MIT Press.

[24] P. Whittle, *Optimization under constraints: theory and applications of nonlinear programming*, Wiley-Inter Science, 1971.

[25] 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.