A CONSTRAINED LINE SEARCH OPTIMIZATION FOR DISCRIMINATIVE TRAINING IN SPEECH RECOGNITION

Cong Liu[1], *Peng Liu*[2], *Hui Jiang*[3], *Frank Soong*[2], *Ren-Hua Wang*[1]

[1]University of Science and Technology of China, Hefei, P. R. China, 230027
[2]Microsoft Research Asia, Beijing, P. R. China, 100080
[3]Department of Computer Science and Engineering, York University, CANADA
Email: yylhbt@mail.ustc.edu.cn pengliu@microsoft.com
hj@cse.yorku.ca frankkps@microsoft.com rhw@ustc.edu.cn

ABSTRACT

In this paper, we propose a novel *constrained line search* to optimize the MMIE objective function for training discriminative HMMs. In our method, the MMI estimation is cast as a constrained maximization problem, where Kullback-Leibler divergence between models before and after parameters adjustment is introduced as a constraint during optimization. Then, based on the idea of line search, we show that a simple, closed-form solution can be derived under some approximation assumptions. The proposed optimization method have been investigated in two speech recognition tasks: TIDIGITS and Switchboard (*mini-train*). Experimental results show that the new training method achieves significant word error rate reduction when comparing with our best MLE models, i.e., relatively 63.8% on TIDIGITS and 6.1% on the Switchboard *mini-train* set, respectively. Our results also show that the *constrained line search* method consistently outperforms the popular EBW method in both tasks.

Index Terms— Discriminative training, Maximum mutual information (MMI), Line search, Kullback-Leibler divergence

1. INTRODUCTION

In the past few decades, discriminative training (DT) has been a very active research topic in the field of speech recognition. Many different discriminative training methods have been proposed to estimate Gaussian mixture continuous density hidden Markov model (CDHMM) in a variety of speech recognition tasks. Estimation of CDHMM parameters is essentially an optimization problem. First of all, we formulate an objective function according to certain estimation criterion, such as maximum mutual information (MMI)[1], minimum classification error (MCE)[2, 3, 4] , minimum word or phone error (MWE or MPE) [5] and minimum divergence (MD)[6]. Secondly, once the objective function is formulated, an effective optimization method must be used to minimize or maximize the objective function w.r.t. all CDHMM parameters. In speech recognition, several different methods have been used to optimize the derived objective function, including GPD (generalized probabilistic descent) algorithm based on first-order gradient descent, the approximate second-order Quickprop method, extended Baum-Welch (EBW) algorithm based on growth transformation and so on. Essentially, all of these optimization methods attempt to search for a nearby local optimal point of the objective function from an initial point according to both a search direction, which is computed based on the first-order derivative (such as gradient), and a step size which is empirically determined in practice. As the result, performance of these optimization methods highly depends on the initial point and property of the objective function. If the derived objective function is highly nonlinear, jagged and non-convex in nature, just like in discriminative training of CDHMMs, it is extremely difficult to optimize it effectively with any simple optimization algorithm.

In this paper, we propose a novel optimization method, called *constrained line search*, to optimize this kind of complicated objective function of Gaussian mixture CDHMMs derived based on the MMI criterion in speech recognition. Firstly, we cast the MMI estimation of CDHMM's as a constrained maximization problem. Under this constraint, the MMI objective function can be approximated by a smoothing quadratic function for each Gaussian means or such like for variances, and the sole optimal point of this function can be easily obtained by vanishing its derivative to zero. Based on this, a closed-form solution to the above constrained maximization for MMIE can be easily derived by a *constrained line search* method. The proposed line search optimization method for MMIE has been investigated for several speech recognition tasks, including connected digit string recognition using TIDIGITS database and large vocabulary recognition using the Switchboard *mini-train* data set. The experimental results clearly show that the proposed line search method outperforms the popular EBW method in both evaluated ASR tasks.

2. A CONSTRAINED OPTIMIZATION FOR MMIE

Assume we have R training utterances $X_1, X_2, \cdots, X_R$ along with their corresponding transcriptions $W_1, W_2, \cdots, W_R$. As we know, the objective function of MMIE takes the following form:

$$\begin{aligned} & \mathcal{F}_{\mathrm{MMI}}(\mathbf{\Lambda} \mid \{X_r, W_r, \mathcal{M}_r\}_{r=1}^{R}, \kappa) \\ = \; & \frac{1}{R} \sum_{r=1}^{R} \log \left[\frac{p^{\kappa}(X_r|\lambda_{W_r}) \cdot p(W_r)}{\sum_{W \in \mathcal{M}_r} p^{\kappa}(X_r|\lambda_W) \cdot p(W)} \right] \end{aligned} \quad (1)$$

where $\mathbf{\Lambda}$ denotes the set of all CDHMMs, and $\mathcal{M}_r$ stands for all competing hypotheses of utterance X_r, and λ_{W_r} denotes composite HMM model for word sequence W_r, and κ $(0 < \kappa \leq 1)$ is the so-called acoustic scaling factor. Usually $\mathcal{M}_r$ is approximately represented by word lattice generated from Viterbi decoding of the utterance X_r.

[1]This work has been done when the first author was a visiting student with Speech Group, Microsoft Research Asia.

Furthermore, we assume that the model set $\mathbf{\Lambda}$ is composed of many individual Gaussian mixture CDHMMs, each of which is represented as $\lambda = (\pi, A, \boldsymbol{\theta})$, where π is initial state distribution, $A = \{a_{ij} \,|\, 1 \le i, j \le N\}$ is transition matrix, and $\boldsymbol{\theta}$ is parameter vector composed of mixture parameters $\theta_i = \{\omega_{ik}, \mu_{ik}, \sigma_{ik}\}_{k=1,2,\cdots,K}$ for each state i, where K stands for the number of Gaussian mixtures in each state. Then, the state observation p.d.f. is assumed to be a mixture of multivariate Gaussian distribution with diagonal covariance matrix:

$$\begin{aligned} p(\mathbf{x}|\theta_i) &= \sum_{k=1}^{K} \omega_{ik} \cdot \mathcal{N}(\mathbf{x}; \mu_{ik}, \sigma_{ik}) \\ &= \sum_{k=1}^{K} \omega_{ik} \prod_{d=1}^{D} \sqrt{\frac{1}{2\pi\sigma_{ikd}^2}}\, \mathrm{e}^{-\frac{(x_d-\mu_{ikd})^2}{2\sigma_{ikd}^2}} \end{aligned} \tag{2}$$

where D is dimension of observation vector $\mathbf{x}$.

In this study, we assume that language model score $p(W)$ is fixed. For any training utterance X and its transcription W, let's consider how to compute acoustic model score $p(X|\lambda_W)$ in the MMIE objective function in eq.(1) based on the composite HMM λ_W of W. Suppose $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T\}$, let $\mathbf{s} = \{s_1, s_2, \cdots, s_T\}$ be the unobserved state sequence, and $\mathbf{l} = \{l_1, l_2, \cdots, l_T\}$ be the associated sequence of the unobserved mixture component labels. Thus, $p(X|\lambda_W)$ is computed as:

$$\begin{aligned} p(X|\lambda_W) \;=\; & \sum_{\mathbf{s}} \sum_{\mathbf{l}} \left\{ \pi_{s_1} \omega_{s_1 l_1} \mathcal{N}(\mathbf{x}_1; \mu_{s_1 l_1}, \sigma_{s_1 l_1}) \right. \\ & \left. \prod_{t=2}^{R} a_{s_{t-1} s_t} \cdot \omega_{s_t l_t} \cdot \mathcal{N}(\mathbf{x}_t; \mu_{s_t l_t}, \sigma_{s_t l_t}) \right\} \end{aligned} \tag{3}$$

where summations are taken over all possible state sequence $\mathbf{s}$ and mixture label $\mathbf{l}$. If we adopt the Viterbi method to approximate the above summation with the single optimal Viterbi path, denoted as $\mathbf{s}^* = \{s_1^*, s_2^*, \cdots, s_T^*\}$ and $\mathbf{l}^* = \{l_1^*, l_2^*, \cdots, l_T^*\}$, then we have

$$\begin{aligned} p(X|\lambda_W) \;\approx\; & \pi_{s_1^*} \omega_{s_1^* l_1^*} \mathcal{N}(\mathbf{x}_1; \mu_{s_1^* l_1^*}, \sigma_{s_1^* l_1^*}) \\ & \prod_{t=2}^{R} a_{s_{t-1}^* s_t^*} \cdot \omega_{s_t^* l_t^*} \cdot \mathcal{N}(\mathbf{x}_t; \mu_{s_t^* l_t^*}, \sigma_{s_t^* l_t^*}) \end{aligned} \tag{4}$$

After substituting eq.(3) or eq.(4) into eq.(1), we can see that the MMI objective function $\mathcal{F}_{\mathrm{MMI}}$ becomes a highly nonlinear complicated function, which is extremely difficult to optimize directly. Thus, we first make the following assumptions: i) assume that all competing hypothesis spaces $\mathcal{M}_r$ are unchanged during optimization; ii) introduce a small scaling factor κ ($\kappa \ll 1$) to smooth the original MMI objective function. Because of this, it is clear that we should explicitly add the constraint that the HMM model parameters $\mathbf{\Lambda}$ can not significantly differ from their initial values $\mathbf{\Lambda}^0$, which is used to generate word lattices $\{\mathcal{M}_r\}$ and to approximate $p(X|\lambda_W)$ in eq.(4), to ensure that all of these assumptions still remain valid during optimization. Obviously, this kind of constraint can be quantitively formulated based on Kullback-Leibler divergence (KLD) between models. Therefore, the MMI training problem of CDHMMs should be formulated as the following constrained maximization problem:

$$\max_{\mathbf{\Lambda}} \;\; \mathcal{F}_{\mathrm{MMI}}(\mathbf{\Lambda} \mid \{X_r, W_r, \mathcal{M}_r\}_{r=1}^{R}, \kappa) \tag{5}$$

$$\text{subject to} \;\; \mathcal{D}(\mathbf{\Lambda}||\mathbf{\Lambda}^{(\mathbf{0})}) \le \rho^2, \tag{6}$$

where $\mathcal{D}(\mathbf{\Lambda}||\mathbf{\Lambda}^{(\mathbf{0})})$ means KL divergence calculated between the model set $\mathbf{\Lambda}$ and its initial value $\mathbf{\Lambda}^{(\mathbf{0})}$, and ρ^2 is a pre-set constant to control the search range.

As we will show later, if the smoothing factor κ is sufficiently small, i.e., $\kappa \ll 1$, under the constraint in eq.(6) the MMI objective function can be approximated as a quadratic function so that a closed-form solution (at least sub-optimal) can be derived for the constrained optimization problem in eqs.(5) and (6).

3. MODEL CONSTRAINTS BASED ON KL DIVERGENCE

Assume the model set $\mathbf{\Lambda}$ is composed of many individual models λ ($\lambda \in \mathbf{\Lambda}$), the KLD constraint in eq.(6) can be equivalently represented for each model as:

$$\mathcal{D}(\lambda \,||\, \lambda^{(0)}) \le \rho^2 \quad (\lambda \in \mathbf{\Lambda}) \tag{7}$$

Furthermore, $\mathcal{D}(\lambda \,||\, \lambda^{(0)})$ can be decomposed according to all Gaussian components:

$$\begin{aligned} \mathcal{D}(\lambda \,||\, \lambda^{(0)}) \;&\le\; \sum_{i=1}^{N} \mathcal{D}(\theta_i \,||\, \theta_i^{(0)}) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} \omega_{ik} \cdot \mathcal{D}(\mathcal{N}(\mu_{ik}, \sigma_{ik}) \,||\, \mathcal{N}(\mu_{ik}^{(0)}, \sigma_{ik}^{(0)})) \end{aligned} \tag{8}$$

It is well known that KL-divergence between two Gaussians can be explicitly computed as:

$$\begin{aligned} & \mathcal{D}(\mathcal{N}(\mu_{ik}, \sigma_{ik}) \,||\, \mathcal{N}(\mu_{ik}^{(0)}, \sigma_{ik}^{(0)})) \\ =\; & \frac{1}{2} \sum_{d=1}^{D} \left[\log \frac{(\sigma_{ikd}^{(0)})^2}{\sigma_{ikd}^2} - 1 + \frac{\sigma_{ikd}^2}{(\sigma_{ikd}^{(0)})^2} + \frac{(\mu_{ikd} - \mu_{ikd}^{(0)})^2}{(\sigma_{ikd}^{(0)})^2} \right] \end{aligned} \tag{9}$$

Furthermore, we decompose the KLD constraint in eq.(9) for Gaussian means μ_{ik} and variances σ_{ik} separately as follows:

$$\sum_{d=1}^{D} \frac{(\mu_{ikd} - \mu_{ikd}^{(0)})^2}{(\sigma_{ikd}^{(0)})^2} \le \rho_1^2 \quad (\text{for all } \lambda \in \mathbf{\Lambda} \text{ and } i, k) \tag{10}$$

$$\sum_{d=1}^{D} \left[\log \frac{(\sigma_{ikd}^{(0)})^2}{\sigma_{ikd}^2} - 1 + \frac{\sigma_{ikd}^2}{(\sigma_{ikd}^{(0)})^2} \right] \le \rho_2^2 \quad (\text{for all } \lambda \in \mathbf{\Lambda} \text{ and } i, k) \tag{11}$$

where ρ_1^2 and ρ_2^2 are two pre-set constants to control constraint range for mean vectors and variance vectors, respectively.

4. CONSTRAINED LINE SEARCH OPTIMIZATION

Now let's consider how to maximize the MMI objective function $\mathcal{F}_{\mathrm{MMI}}$ under the constraints in eqs.(10) and (11). First of all, we consider the first-order derivative of $\mathcal{F}_{\mathrm{MMI}}$ with respect to a Gaussian mean vector μ_{ik} as follows:

$$\begin{aligned} \frac{\partial \mathcal{F}_{\mathrm{MMI}}}{\partial \mu_{ik}} \;&=\; \frac{1}{R} \sum_{r=1}^{R} \sum_{t=1}^{T} \left[\gamma_{ik}^{\mathrm{num}}(r,t) - \gamma_{ik}^{\mathrm{den}}(r,t) \right] \cdot c_{ik}(r,t) \\ & \qquad \cdot \frac{\partial}{\partial \mu_{ik}} \log \left[\omega_{ik} \mathcal{N}(x_{rt}; \mu_{ik}, \sigma_{ik}) \right] \\ &= \frac{1}{R} \sum_{r=1}^{R} \sum_{t=1}^{T} \left[\gamma_{ik}^{\mathrm{num}}(r,t) - \gamma_{ik}^{\mathrm{den}}(r,t) \right] \cdot c_{ik}(r,t) \cdot \frac{(x_{rt} - \mu_{ik})}{\sigma_{ik}^2} \end{aligned} \tag{12}$$

where $\gamma_{ik}^{\mathrm{num}}(r,t)$ and $\gamma_{ik}^{\mathrm{den}}(r,t)$ are occupancy probability of r-th training data X_r at time t for Gaussian k of state i collected based on numerator lattice and denominator lattice respectively, and $c_{ik}(r,t)$ denotes occupancy probability for k-th Gaussian in state i:

$$c_{ik}(r,t) = \frac{\omega_{ik} \cdot \mathcal{N}(x_{rt}; \mu_{ik}, \sigma_{ik})}{\sum_{l=1}^{K} \omega_{sl} \cdot \mathcal{N}(x_{rt}; \mu_{ik}, \sigma_{il})}. \tag{13}$$

Obviously, $\gamma_{ik}^{\mathrm{num}}(r,t)$, $\gamma_{ik}^{\mathrm{den}}(r,t)$ and $c_{ik}(r,t)$ are all functions of model parameter $\mathbf{\Lambda}$. However, if the scaling factor κ is sufficiently small and the model parameter $\mathbf{\Lambda}$ is limited by the constraint eq.(6), they are all actually slowly-changing w.r.t. $\mathbf{\Lambda}$ and can be approximately considered as constants w.r.t $\mathbf{\Lambda}$. Then, under these conditions, from eq.(12) we can see that $\mathcal{F}_{\mathrm{MMI}}$ is approximately a quadratic function of μ_{ik}. Its sole optimal point $\hat{\mu}_{ik}$ can be easily computed by vanishing its derivative to zero $\frac{\partial \mathcal{F}_{\mathrm{MMI}}}{\partial \mu_{ik}} = 0$ as:

$$\hat{\mu}_{ik} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T}\left[\gamma_{ik}^{\mathrm{num}}(r,t) - \gamma_{ik}^{\mathrm{den}}(r,t)\right] \cdot \frac{c_{ik}(r,t)}{\sigma_{ik}^2} \cdot x_{rt}}{\sum_{r=1}^{R}\sum_{t=1}^{T}\left[\gamma_{ik}^{\mathrm{num}}(r,t) - \gamma_{ik}^{\mathrm{den}}(r,t)\right] \cdot \frac{c_{ik}(r,t)}{\sigma_{ik}^2}} \tag{14}$$

However, since the MMI objective function $\mathcal{F}_{\mathrm{MMI}}$ is indefinite, the above optimal point may be maximum for some μ_{ik} but minimum for others, as shown in Figure 1. In total, we have four possible situations: i) $\hat{\mu}_{ik}$ is maximum and it is located within constraint range, as in case 1; ii) $\hat{\mu}_{ik}$ is maximum but it is outside constraint range, as in case 2; iii) $\hat{\mu}_{ik}$ is minimum, as in case 3; iv) $\mathcal{F}_{\mathrm{MMI}}$ degenerates to a linear function when $\gamma_{ik}^{\mathrm{num}}(r,t) = \gamma_{ik}^{\mathrm{den}}(r,t)$, no optimal exists as shown in case 4. Among these cases, even when $\hat{\mu}_{ik}$ is indeed a maximum, it may not be a good solution to eqs.(5) and (6) since it may be very far from the initial point so that the constraint in eq.(6) is not satisfied, as in case 2.

Therefore, in this paper, for the objective function $\mathcal{F}_{\mathrm{MMI}}$ whose optimal point can be obtained (case 1, 2, 3), we propose to conduct a *constrained line search* along the line joining the initial $\mu_{ik}^{(0)}$ and the optimal $\hat{\mu}_{ik}$ to search for an optimal point which maximizes the objective function $\mathcal{F}_{\mathrm{MMI}}$ subject to the constraint in eq.(10):

$$\mu_{ik}^{*} = (1 - \epsilon_{ik}) \cdot \mu_{ik}^{(0)} + \epsilon_{ik} \cdot \hat{\mu}_{ik} \quad (-\infty < \epsilon_{ik} < \infty) \tag{15}$$

where ϵ_{ik} is a linear interpolation weight, which is determined by maximizing $\mathcal{F}_{\mathrm{MMI}}$ w.r.t. μ_{ik} under the constraint eq.(10). Due to the simple format of constraint eq.(10), ϵ_{ik} can be easily computed for all of these three different cases. After defining the KL divergence for mean vectors in eq.(10) as $\mathcal{D}(\mu_{ik}||\mu_{ik}^{(0)}) = \sum_{d=1}^{D} \frac{(\mu_{ikd} - \mu_{ikd}^{(0)})^2}{(\sigma_{ikd}^{(0)})^2}$, the general form to compute ϵ_{ik} for eq.(15) is given as follows, for the above different cases:

$$\epsilon_{ik} = \begin{cases} 1 & \text{if case 1} \\ \sqrt{\frac{\rho_1^2}{\mathcal{D}(\hat{\mu}_{ik}||\mu_{ik}^{(0)})}} & \text{if case 2} \\ -\sqrt{\frac{\rho_1^2}{\mathcal{D}(\hat{\mu}_{ik}||\mu_{ik}^{(0)})}} & \text{if case 3} \end{cases} \tag{16}$$

For the linear case in which no optimal point exists (case 4), we need to employ the gradient $\nabla\mathcal{F}_{\mathrm{MMI}}$ as the direction of search instead of that of line search, which can be expressed as follows:

$$\mu_{ik}^{*} = \mu_{ik}^{(0)} + \epsilon_{ik} \cdot \nabla\mathcal{F}_{\mathrm{MMI}}(\mu_{ik}^{(0)}) \quad (0 < \epsilon_{ik} < \infty) \tag{17}$$

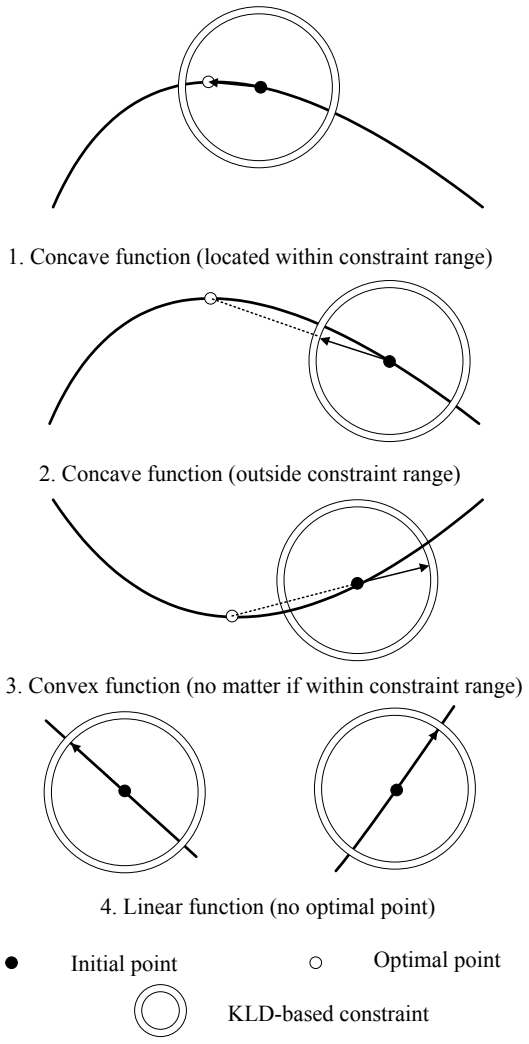

Fig. 1. Illustration of Constrained Line Search for maximizing the objective function in several cases.

after substituting eq.(17) into eq.(10), the interpolation weight ϵ_{ik} for linear case can be obtained as follows:

$$\epsilon_{ik} = \sqrt{\frac{\rho_1^2}{\sum_{d=1}^{D}\left[\nabla\mathcal{F}_{\mathrm{MMI}}(\mu_{ikd}^{(0)})/(\sigma_{ikd}^{(0)})^2\right]}} \quad \text{if case 4} \tag{18}$$

Similarly, other model parameters, such as Gaussian variances, mixture weights, etc., can be optimized following the same idea of constrained line search. Due to space limit, we have to leave out all formula for updating other HMM parameters in this paper.

In summary, during the proposed MMIE training process, we first generate word-lattices for all training data based on an initial model. Then all statistics are collected from these word-lattice based on the initial model and all Gaussian means in the model set are updated according to eq.(15). Next, we can re-collect statistics based on the updated model and update models again until it converges. In this work, for simplicity, we only update Gaussian means in the MMIE training with the proposed constrained line search method.

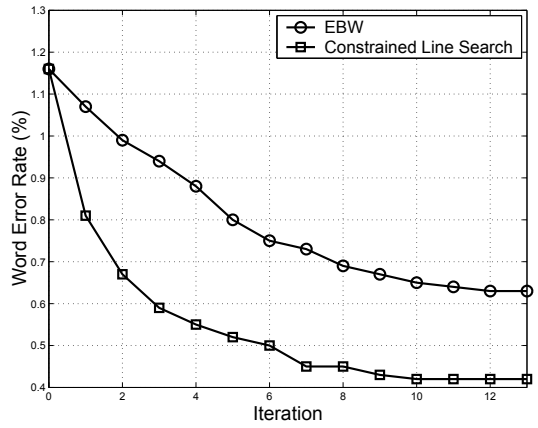

Fig. 2. Comparison of word error rates (in %) of different optimization methods on the TIDIGITS test set.

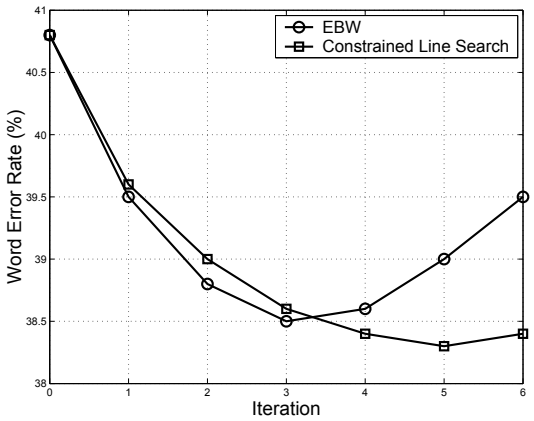

Fig. 3. Comparison of word error rates (in %) of different optimization methods on the Switchboard *eval2000* test set.

5. EXPERIMENTS

The proposed constrained line search optimization method for the MMIE training has been evaluated on two speech recognition tasks: connected digit string recognition by uisng the TIDIGITS database and large vocabulary speech recognition task by using the Switchboard *mini-train* data set. In TIDIGITS, 39-dimensional MFCCs (12-d static MFCC, log-energy, delta and acceleration coefficients) are used to train 10-state whole-word based models, with 6 mixtures per state. And there are 12,549 utterances in the training set and 12,547 utterances in the testing set. In Switchboard *mini-train* task, 39-dimensional PLPs are used to train context dependent tri-phone HMM, with 12 mixtures per state. The training set has totally 18,252 utterances, and we use Switchboard *eval2000* data set (1,831 utterances) as the test set.

In our experiments, we always use the best ML-trained HMMs as the seed model for the MMIE training on both databases. The MMIE training is conducted by using two different optimization methods, namely the popular EBW method and the proposed constrained line search method. Please note that we update all Gaussian means, variances and mixture weights in the EBW method while we update only Gaussian means in the proposed constrained line search method. For the EBW method, the results by using the method in [5] to determine the constant D ($E = 2$) in each iteration are represented. In the constrained line search method, the constant ρ^2 to control search range is decreased with iterations to make training more stable and convergence faster.

In Fig.2, we give a performance comparison between the proposed line search method and the popular EBW method on the test set of TIDIGITS. The results clearly show that the proposed method achieves larger improvement than the EBW method. In the constrained line search, the word error rate decreases from 1.16% to 0.42%, which indicates 63.8% relative error reduction, from the best MLE-trained models. The learning curves in Fig.2 show that the constrained line search method converges pretty well.

In Fig.3, we also give the word error performance comparison on Switchboard *eval2000* data set by using different optimization methods for the MMIE training. It shows that the proposed line search method also outperforms the EBW method. For the constrained line search, the word error rate decreases from 40.8% to 38.3%, or a 6.1% relative error reduction in comparing with our best MLE-trained model.

6. CONCLUSIONS AND FUTURE WORK

In this paper, a new *constrained line search* method is proposed to optimize the MMIE objective function for discriminative training. Experimental results show that our method consistently achieves better performance than the popular EBW method in two speech recognition tasks.

In this work, we only update means for simplicity. More experiments to update other HMM parameters with the same line search idea are underway. Furthermore, the same line search idea can also be applied to other discriminative training criteria, such as MCE, MPE, MD, etc. These works will be reported in the future.

7. REFERENCES

[1] 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.

[2] 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.

[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," *IEEE Trans. on Speech and Audio Processing,* pp.945-955, Vol. 13, No.5, September 2005.

[4] B. Liu, H. Jiang, J.-L. Zhou, R.-H. Wang, "Discriminative Training Based on The Criterion of Least Phone Competing Tokens for Large vocabulary Speech Recognition," Proc. of 2005 IEEE International Conference on Acoustic, Speech, Signal Processing (ICASSP'2005), Philadelphia, Mar. 2005.

[5] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," *Proc. of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP'2002)*, Orlando, 2002.

[6] J. Du, P. Liu, F.K. Soong, J.-L. Zhou, R.-H Wang, "Minimum Divergence based Discriminative Training", *Proc. ICSLP*, pp. 2410-2413 , 2006.