A TRUST REGION BASED OPTIMIZATION FOR MAXIMUM MUTUAL INFORMATION ESTIMATION OF HMMS IN SPEECH RECOGNITION

Zhi-Jie Yan[*1] *Cong Liu*[1] *Yu Hu*[1] *Hui Jiang*[2]

[1]iFlytek Speech Lab, University of Science and Technology of China, Hefei, P. R. China
[2]Department of Computer Science and Engineering, York University, Toronto, Canada
yanzhijie@ustc.edu *yylhbt@mail.ustc.edu.cn* *yuhu@iflytek.com* *hj@cse.yorku.ca*

ABSTRACT

In this paper, we present a new optimization method for MMIE-based discriminative training of HMMs in speech recognition. In our method, the MMIE training of Gaussian mixture HMMs is formulated as a so-called trust region problem, where a quadratic objective function is minimized under a spherical constraint, so that an efficient global optimization method for the trust region problem can be used to solve the MMIE training problem of HMMs. Experimental results on the WSJ0 Nov'92 evaluation task demonstrate that the trust region based optimization significantly outperforms the conventional EBW method in terms of optimization convergence behavior as well as speech recognition performance. It has been observed that the trust region method achieves up to 23.3% relative recognition error reduction over a well-trained MLE system while the EBW method gives only 13.3% relative error reduction.

Index Terms— Speech recognition, Hidden Markov models, Optimization methods

1. INTRODUCTION

Recently, discriminative training (DT) methods have achieved a tremendous success in a variety of speech recognition tasks. Many different DT methods have been proposed to estimate Gaussian mixture continuous density hidden Markov models (CDHMMs). Discriminative training 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), minimum classification error (MCE), minimum word or phone error (MWE or MPE), etc. Secondly, an effective optimization method is 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 etc.

In this paper, we develop a trust region based HMM parameter optimization method for discriminative training of HMMs. Following an approximation-maximization manner, we first derive an auxiliary function to approximate the original MMI objective function of HMMs in a close neighborhood of initial model parameters. For Gaussian mixture CDHMMs, the auxiliary function is a quadratic function. Next, we propose to impose a locality constraint for all Gaussian kernels to ensure that the above approximation remains valid during optimization process. Under some conditions, the locality constraint can be relaxed as a spherical or elliptic constraint. As a result, the MMIE training of HMMs can be formulated as a trust region problem where a quadratic objective function is minimized under a spherical or elliptic constraint. Since there exists an efficient algorithm to find the global optimum for this type of trust region problem, it can be used to solve the MMIE training of HMMs in a fairly efficient way. Compared with conventional EBW method, the trust region based optimization is a bounded optimization method so that the stability and generalization ability of parameter optimization can thus be improved. Since the global optimal point of the auxiliary function is guaranteed to be found in each iteration, the trust region method converges faster than other optimization methods.

In this work, we have evaluated our trust region based optimization method for the MMIE-based discriminative training in speech recognition on the standard WSJ0 Nov'92 evaluation task. Experimental results show that the trust region based optimization significantly outperforms the conventional EBW method in terms of convergence behavior as well as speech recognition performance.

The rest of this paper is organized as follows: in Section 2, the trust region problem in mathematic literature is briefly reviewed; in Section 3 and 4, the MMIE training problem is formulated as a trust region based optimization problem; Section 5 gives our experimental results; and finally in Section 6, we will draw our conclusions.

2. TRUST REGION PROBLEM AND ITS SOLUTION

We know that most non-convex optimization problems are difficult to solve. One of few non-convex optimization problems with efficient algorithm is minimization of a quadratic function under a sphere or elliptic constraint. This problem arises as a special case in a number of nonlinear programming problems, which are usually called *trust region (TR)* problems [1]. In this section, we briefly review the optimization theory to show how the *global* minimum of this TR problem can be efficiently found by a fast algorithm.

Recall that a general quadratic function of a n-variable vector, $\mathbf{u}$, has the form $\frac{1}{2}\mathbf{u}^\top Q\mathbf{u}+\mathbf{q}^\top\mathbf{u}$, where Q is a symmetric matrix and $\mathbf{q}$ a vector. The TR problem is expressed as:

$$\min_{\mathbf{u}\in\mathbf{R}^n} \quad \frac{1}{2}\mathbf{u}^\top Q\mathbf{u}+\mathbf{q}^\top\mathbf{u} \qquad \text{s.t.} \quad \mathbf{u}^\top\mathbf{u}\le\rho^2, \tag{1}$$

with ρ a constant to control size of the spherical trust region. If Q is positive definite, a global minimum to Eq. (1) can be calculated as $\hat{\mathbf{u}}=-Q^{-1}\mathbf{q}$. Furthermore, if the norm of $\hat{\mathbf{u}}$ is bounded by ρ^2, i.e., $\hat{\mathbf{u}}^\top\hat{\mathbf{u}}\le\rho^2$, then $\hat{\mathbf{u}}$ is a feasible solution of the TR problem in Eq. (1). In all other cases, the global minimum of Eq. (1) can also be found efficiently according to the following theorem [1]:

*This author is now with the speech group of Microsoft Research Asia.

Theorem 1 *The vector $\mathbf{u}^*$ is the global solution to the trust region problem in Eq. (1) if and only if $\mathbf{u}^*$ is feasible and there is a scalar $\lambda \geq 0$ such that the following conditions are satisfied:*

$$\begin{aligned} (Q + \lambda I)\mathbf{u}^* &= -\mathbf{q}, \\ \lambda(\mathbf{u}^{*\top}\mathbf{u}^* - \rho^2) &= 0, \\ (Q + \lambda I) &\text{ is positive semi-definite,} \end{aligned} \tag{2}$$

where I denotes identity matrix.

As proven in [1], the conditions in Eq. (2) are both necessary and sufficient conditions of that $\mathbf{u}^*$ is globally minimum of Eq. (1). Based on the first condition in Eq. (2), the global minimum $\mathbf{u}^*$ can be easily calculated based on a scalar λ as:

$$\mathbf{u}^* = -(Q + \lambda I)^{-1}\mathbf{q}. \tag{3}$$

Therefore, the TR problem in Eq. (1) turns out into a much easier problem to search for a scalar λ to satisfy that $(Q + \lambda I)$ is positive semi-definite and the norm of $\mathbf{u}^*$ equal to ρ^2, i.e. $\mathbf{u}^{*\top}\mathbf{u}^* = \|(Q + \lambda I)^{-1}\mathbf{q}\|_2 = \rho^2$.

Moreover, another theorem in [1] is useful for searching the scalar λ for $\mathbf{u}^*$. Define λ_0 as the minimum λ such that $Q + \lambda I$ is positive semi-definite. And it is easy to see that λ_0 is the negative of the smallest (closest to -∞) eigenvalue of Q.

Theorem 2 *If $\mathbf{q} \neq 0$, λ_1 and λ_2 are two scalars that satisfy $\lambda_0 \leq \lambda_1 < \lambda_2$. Let $\mathbf{u}_1^*$ and $\mathbf{u}_2^*$ are solutions to $(Q + \lambda_1 I)\mathbf{u}_1^* = -\mathbf{q}$ and $(Q + \lambda_2 I)\mathbf{u}_2^* = -\mathbf{q}$ respectively, then $\|\mathbf{u}_1^*\|_2 > \|\mathbf{u}_2^*\|_2$.*

In other words, the norm $\|(Q + \lambda I)^{-1}\mathbf{q}\|_2$ is a monotonic decreasing function of λ for $\lambda > \lambda_0$. As a result, the unique scalar, denoted as λ^*, which satisfies $\|(Q + \lambda^* I)^{-1}\mathbf{q}\|_2 = \rho^2$, can be efficiently found in the interval $[\lambda_0, \infty)$ using a binary search method.

3. MMIE AS CONSTRAINED OPTIMIZATION

For a training set containing R utterances $\{O_1, \cdots, O_R\}$, the MMIE objective function can be written as:

$$\begin{aligned} \mathbf{F}_{\text{MMI}}(\Lambda) &= \frac{1}{R}\sum_r \ \mathcal{F}_r(\Lambda|O_r) \\ &= \frac{1}{R}\sum_r \Big[\log p(O_r \mid \Lambda, \mathcal{M}_r^+) - \log p(O_r \mid \Lambda, \mathcal{M}_r^-)\Big] \end{aligned} \tag{4}$$

where Λ represents the set of all HMM parameters, $\mathcal{M}_r^+$ and $\mathcal{M}_r^-$ stand for the reference model space and competing model space of O_r, respectively. In MMIE training, $\mathcal{M}_r^+ = \{W_r\}$, which is the reference word sequence, while $\mathcal{M}_r^-$ is composed of all possible word sequences, which are usually represented by a word lattice or graph.

As shown in [2], during the optimization process of the above MMIE objective function, it is beneficial to impose a local constraint on model parameters Λ to ensure that they do not deviate too much from its initial values, i.e., $\Lambda^{(n)}$. The local constraint can be quantitatively defined based on Kullback-Leibler divergence (KLD). Therefore, MMIE training of HMM parameters, Λ, can be formulated as the following iterative constrained maximization problem:

$$\Lambda^{(n+1)} = \arg\max_{\Lambda} \ \mathbf{F}_{\text{MMI}}(\Lambda) \tag{5}$$

$$\text{subject to} \quad \mathcal{D}(\Lambda||\Lambda^{(n)}) \leq \rho^2, \tag{6}$$

where $\mathcal{D}(\Lambda||\Lambda^{(n)})$ is the KLD between Λ and $\Lambda^{(n)}$, and $\rho > 0$ is a pre-set constant to control the search range.

4. FORMULATING MMIE AS TRUST REGION PROBLEM

In the following, we consider to convert the above constrained optimization of MMIE into a trust region (TR) problem in Eq. (1) so that it can efficiently solved using the fast algorithm introduced in Section 2.

4.1. Transformation of Locality Constraint

Assume that there are totally $\mathcal{K}$ Gaussian mixtures in the CDHMM set, i.e., $\Lambda = \{\lambda_k \mid k = 1, \cdots, \mathcal{K}\}$, where λ_k denotes a multivariate Gaussian distribution with mean vector μ_k and covariance matrix Σ_k, i.e., $\mathcal{N}(\mu_k, \Sigma_k)$ where $k \in (1, 2, \ldots, \mathcal{K})$.

As shown in [2], the KLD-based constraint in Eq. (6) can be relaxed as sum of all individual Gaussians as follows:

$$\mathcal{D}(\Lambda||\Lambda^{(n)}) \leq \sum_k \mathcal{D}(\lambda_k \,||\, \lambda_k^{(n)}) \leq \frac{\rho^2}{2}. \tag{7}$$

The KLD for each Gaussian, $\mathcal{D}(\lambda_k \,||\, \lambda_k^{(n)})$, can be calculated by the following closed-form formula as:

$$\begin{aligned} \mathcal{D}(\lambda_k \parallel \lambda_k^{(n)}) = \frac{1}{2}\Big[&(\mu_k - \mu_k^{(n)})^\top \Sigma_k^{-1}(\mu_k - \mu_k^{(n)}) \\ &+ \text{tr}\big(\Sigma_k \Sigma_k^{(n)^{-1}}\big) + \log\frac{|\Sigma_k^{(n)}|}{|\Sigma_k|} - M\Big] \end{aligned} \tag{8}$$

where M is dimension of observation vectors.

If we only consider to optimize mean vectors, μ_k, of HMMs and assume covariance matrices, Σ_k, are constant during the MMIE training, the locality constraint in Eq. (7) can be simplified as:

$$\sum_k (\mu_k - \mu_k^{(n)})^\top \Sigma_k^{(n)-1}(\mu_k - \mu_k^{(n)}) \leq \rho^2. \tag{9}$$

We first normalize each mean vector with its corresponding covariance matrix as $\hat{\mu}_k = \Sigma_k^{(n)-\frac{1}{2}}(\mu_k - \mu_k^{(n)})$, then we concatenate all normalized mean vectors as a large single super-vector:

$$\mathbf{u} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \vdots \\ \hat{\mu}_{\mathcal{K}} \end{bmatrix}_{(D\mathcal{K}\times 1)}. \tag{10}$$

Finally, the locality constraint in Eq. (9) can be rewritten as a spherical constraint:

$$\mathbf{u}^\top \mathbf{u} \leq \rho^2. \tag{11}$$

4.2. Approximation of MMIE Objective Function

Based on the expectation-based approximation (E-approx) method in [3], the log-likelihood function of HMMs can be approximated by an auxiliary function $\mathcal{Q}(\Lambda|\Lambda^{(n)})$ as follows:

$$\begin{aligned} &\log p(O_r \mid \Lambda, \mathcal{M}_r) \approx \mathcal{Q}_r(\Lambda|\Lambda^{(n)}) \\ &\equiv \sum_{\mathbf{l}\in\mathcal{M}_r}\sum_{\mathbf{s_l}} \Big[\log p(O_r, \mathbf{l}, \mathbf{s}_l|\Lambda) \cdot \Pr(\mathbf{l}, \mathbf{s_l}|O_r, \Lambda^{(n)})\Big] - \mathcal{H}(\Lambda^{(n)}) \end{aligned} \tag{12}$$

where $\mathbf{l}$ denotes a complete path in $\mathcal{M}_r$ and $\mathbf{s_l}$ are all missing data along the path $\mathbf{l}$, and $\mathcal{H}(\Lambda^{(n)}) \equiv \sum_{\mathbf{l}\in\mathcal{M}_r}\sum_{\mathbf{s_l}}[\log \Pr(\mathbf{l}, \mathbf{s_l}|O_r, \Lambda^{(n)})$.

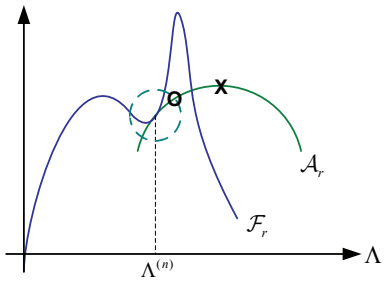

Fig. 1. Trust region constraint (dashed circle) for parameter optimization. 'O' is the optimal of $\mathcal{A}_r$ with constraint, while 'X' is the optimal without constraint.

$\Pr(\mathbf{l}, \mathbf{s_l}|O_r, \Lambda^{(n)})]$ stands for entropy of missing data $\mathbf{l}$ and $\mathbf{s_l}$ calculated with $\Lambda^{(n)}$.

Based on the E-approx in Eq. (12), each term $\mathcal{F}_r(\Lambda|O_r)$ in Eq. (4) can be approximated by the following auxiliary function $\mathcal{A}_r$ as:

$$\mathcal{A}_r(\Lambda|\Lambda^{(n)}) = \left[\mathcal{Q}_r^+(\Lambda \mid \Lambda^{(n)}) - \mathcal{Q}_r^-(\Lambda \mid \Lambda^{(n)})\right] + C, \quad (13)$$

where $\mathcal{Q}_r^+$ is calculated as in Eq. (12) based on $\mathcal{M}_r^+$, $\mathcal{Q}_r^-$ is computed based on $\mathcal{M}_r^-$, and C is a constant independent of Λ.

Based on the discussions in [3], it is straightforward to prove that:

$$\mathcal{F}_r(\Lambda|O_r)\Big|_{\Lambda=\Lambda^{(n)}} = \mathcal{A}_r(\Lambda|\Lambda^{(n)})\Big|_{\Lambda=\Lambda^{(n)}} \quad (14)$$

$$\frac{\partial \mathcal{F}_r(\Lambda|O_r)}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(n)}} = \frac{\partial \mathcal{A}_r(\Lambda|\Lambda^{(n)})}{\partial \Lambda}\Big|_{\Lambda=\Lambda^{(n)}} \quad (15)$$

Obviously, $\mathcal{A}_r$ can be viewed as a local approximation of $\mathcal{F}_r$ around the initial model point $\Lambda^{(n)}$ with accuracy up to the first order derivative. Therefore, if we optimize the auxiliary function $\mathcal{A}_r$ subject to the locality constraint, it may indirectly improve the original objective function $\mathcal{F}_r$ as well. It should be noted, however, that optimization of the auxiliary function $\mathcal{A}_r$ does not necessarily lead to the optimal solution to the objective function $\mathcal{F}_r$ since $\mathcal{A}_r$ is only a local approximation of $\mathcal{F}_r$ around the initial model point $\Lambda^{(n)}$. Therefore, a locality constraint as in Eq. (7) is necessary to ensure model parameters will not deviate too much from their initial values so that $\mathcal{A}_r$ always serves as a good approximation of $\mathcal{F}_r$. Fig. 1 illustrates the constraint (dashed circle) imposed for optimization. When a constraint with appropriate size is used, optimizing $\mathcal{A}_r$ will improve $\mathcal{F}_r$ as well.

From Eqs. (12) and (13), the auxiliary function $\mathcal{A}_r(\cdot)$ can be computed as:

$$\begin{aligned}\mathcal{A}_r(\Lambda|\Lambda^{(n)}) &= \sum_{\mathbf{l}\in\mathcal{M}_r^+}\sum_{\mathbf{s_l}}\left[\Pr(\mathbf{l}, \mathbf{s_l} \mid O_r, \Lambda^{(n)}) \cdot \log p(O_r, \mathbf{l}, \mathbf{s_l} \mid \Lambda)\right] \\ &- \sum_{\mathbf{l}\in\mathcal{M}_r^-}\sum_{\mathbf{s_l}}\left[\Pr(\mathbf{l}, \mathbf{s_l} \mid O_r, \Lambda^{(n)}) \cdot \log p(O_r, \mathbf{l}, \mathbf{s_l} \mid \Lambda)\right] + C \\ &= \sum_k\sum_t \left(\gamma_{krt}^+ - \gamma_{krt}^-\right) \log p(O_{rt}, k \mid \Lambda) + C' \quad (16)\end{aligned}$$

where γ_{krt}^+ and γ_{krt}^- denote occupancy statistics for k-th Gaussian kernel collected based on $\mathcal{M}_r^+$ and $\mathcal{M}_r^-$ respectively, and C' is constant independent of Λ. If we only optimize mean vectors of HMMs, we have:

$$\log p(O_{rt}, k \mid \Lambda) = -\frac{1}{2}(O_{rt}-\mu_k)^\top \Sigma_k^{(n)-1}(O_{rt}-\mu_k) + c_{rk} \quad (17)$$

where c_{rk} is another constant independent of Gaussian mean vectors. Then if we define

$$\xi_k = \sum_r\sum_t(\gamma_{krt}^+ - \gamma_{krt}^-) \quad (18)$$

$$g_k = \sum_r\sum_t(\gamma_{krt}^+ - \gamma_{krt}^-)\Sigma_k^{(n)-\frac{1}{2}}(\mu_k^{(n)} - O_{rt}) \quad (19)$$

and construct a matrix and vector as:

$$Q = \begin{bmatrix} \xi_1 \cdot I_{D\times D} & & & \\ & \xi_2 \cdot I_{D\times D} & & \\ & & \ddots & \\ & & & \xi_{\mathcal{K}} \cdot I_{D\times D} \end{bmatrix} \quad (20)$$

$$\mathbf{q} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_{\mathcal{K}} \end{bmatrix}_{(D\mathcal{K}\times 1)}. \quad (21)$$

The MMIE training of HMMs in Eq. (4) can be converted to the TR problem as:

$$\begin{aligned}\Lambda^* &= \arg\min_\Lambda \ \mathbf{F}_{\rm MMI}(\Lambda) \approx \arg\min_\Lambda \ \sum_r \mathcal{A}_r(\Lambda|\Lambda^{(n)}) \\ &\equiv \arg\min_\Lambda \ \left[\frac{1}{2}\mathbf{u}^\top Q\mathbf{u} + \mathbf{q}^\top\mathbf{u}\right] \quad (22)\end{aligned}$$

which subject to the locality constraint in Eq. (11), i.e. $\mathbf{u}^\top\mathbf{u} \le \rho^2$.

Obviously, the efficient global optimization method in Section 2 can be used to solve the TR problem in Eq. (22) for MMIE training of HMMs. In this case, the Q matrix is a block diagonal matrix which is not positive definite. So we need to search for a λ that makes $(Q + \lambda I)$ positive definite, and ensure that the solution $\mathbf{u} = -(Q + \lambda I)^{-1}\mathbf{q}$ satisfies the constraint $\mathbf{u}^\top\mathbf{u} = \rho^2$. According to Theorem 2, this problem can be solved efficiently with a binary search of λ. In our experiments, the computational cost spent for finding the optimal λ^* is negligible when compared with the cost to collect statistics to compute Q and $\mathbf{q}$ from the whole training set.

Note that optimization of Gaussian covariance matrices can also be similarly formulated as a trust region problem based on the second order Taylor series approximation. Due to space limit, we will report those results somewhere else in the future.

5. EXPERIMENTS

We have evaluated the above trust region based optimization method for MMIE training of HMMs in speech recognition on the WSJ0 database. Our training set is the standard SI-84 set, consisting of 7,133 utterances from 84 speakers. Evaluation is performed on the standard Nov'92 non-verbalized 5k close-vocabulary test set (wsj-5k), including 330 utterances from 8 speakers. For the MLE baseline, we use a similar setup as the WSJ HTK recipe in [4, 5]. Cross-word tri-phone HMMs with a total number of 2,774 tied-states are trained, and each state has 8 Gaussian components. The word error rate (WER) of the MLE baseline using standard tri-gram language model is 4.89%. This result is comparable with the best results on this task reported in [4, 5].

For MMIE discriminative training, we evaluate and compare two parameter optimization methods: one is the conventional extended Baum-Welch (EBW), which is implemented using the latest release

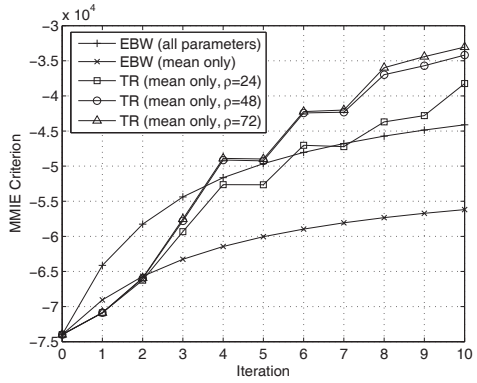

Fig. 2. Optimization of MMIE criterion using EBW method and trust region based method with different region size ρ.

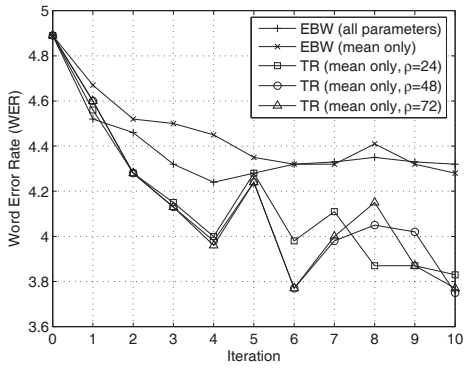

Fig. 3. WER reduction using EBW method and trust region based method with different region size ρ.

of HTK [6]; the other one is the trust region based parameter optimization proposed in this paper. For trust region based parameter optimization, we only optimize mean vectors of the HMMs. For EBW, we optimize mean vectors only in the first set of experiments, and simultaneously update all parameters, including means, covariances and mixture weights, in the second set of experiments. All other training parameters are set to typical values suggested by HTK Book [6] or based on our previous experiments on the wsj-5k task [7], e.g., the learning constant $E = 2$, i-smoothing $\tau = 100$ for EBW; acoustic scaling factor $\kappa = 1/15$.

5.1. Performance Comparison in terms of MMIE Criterion

Firstly, we compare performance of the two optimization methods in terms of improving the MMIE objective function. Fig. 2 shows the learning curves for EBW and TR (with $\rho = 24/48/72$). The results show that the EBW method improves the MMIE criterion more effectively at the first few iterations while the TR method significantly outperforms EBW at the later iterations. This is mainly due to the fact that the EBW method is an unconstrained optimization, while the TR method constrains model parameters within a trust region that is considered to be trustworthy. The TR method finally converges to a better solution because a global optimum of the auxiliary function is always found at each iteration.

5.2. Performance Comparison in terms of WER

Secondly, we compare speech recognition performance between EBW and TR in terms of WER at each iteration. The results in Fig. 3 show that the TR methods (updating means only) with various region sizes consistently outperform the EBW methods (either updating means only or updating all parameters). The best result using the EBW method is 4.24% in WER (13.3% relative error reduction over MLE). On the other hand, the best recognition performance with the TR training is 3.75% in WER, corresponding to 23.3% relative error reduction over the MLE. To our knowledge, this result is one of the best results reported on this task without multiple systems combination.

Besides, it is also observed that the trust region size ρ directly affects the convergence behavior of the algorithm. A large ρ value typically gives better convergence of the objective function, as shown in Fig. 2. But a too large ρ value may cause fluctuation in WER, as shown in Fig. 3. The best recognition result is achieved using $\rho = 48$ in this experiment. We suggest to set $\rho^2 = 0.02 \sim 0.1 \times$ # Gaussian kernels for the trust region based HMM parameter optimization.

6. CONCLUSIONS

This paper presents a trust region based parameter optimization method for MMIE-based discriminative training of HMMs in speech recognition. This method derives an auxiliary function to approximate the original objective function, and imposes a locality constraint to ensure the auxiliary function serves as a good approximation of the objective function during optimization. The trust region based optimization can be solved effectively by using a fast global optimization algorithm proposed in optimization theory. Experimental results on the WSJ0 Nov'92 5k task show that the proposed trust region method yields better performance than the conventional EBW method, in terms of both criterion improvement and recognition WER reduction.

7. REFERENCES

[1] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, 2nd edition, 2006.

[2] P. Liu, C. Liu, H. Jiang, F. Soong, and R.-H. Wang, "A Constrained Line Search Optimization Method for Discriminative Training of HMMs," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 900–909, 2008.

[3] Hui Jiang and Xinwei Li, *A General Approximation-Optimization Approach to Large Margin Estimation of HMMs*, chapter 7, pp. 103–120, Robust Speech Recognition and Understanding. I-TECH Education and Publishing, 2007.

[4] K. Vertanen, "HTK Wall Street Journal (WSJ) training recipe," http://www.inference.phy.cam.ac.uk/kv227/htk/.

[5] P. Woodland, J. Odell, V. Valtchev, and S. Young, "Large vocabulary continuous speech recognition using HTK," in *Proc. ICASSP1994*, 1994, vol. 2, pp. 125–128.

[6] S. Young, et al., *The HTK Book*, Cambridge University, 2006, Revised for HTK version 3.4.

[7] Z.-J. Yan, B. Zhu, Y. Hu, and R.-H. Wang, "Minimum Word Classification Error Training of HMMs for Automatic Speech Recognition," in *Proc. ICASSP2008*, 2008, pp. 4521–4524.