# Trust Region-Based Optimization for Maximum Mutual Information Estimation of HMMs in Speech Recognition

Cong Liu, Yu Hu, Li-Rong Dai, and Hui Jiang, *Senior Member, IEEE*

***Abstract*—In this paper, we have proposed two novel optimization methods for discriminative training (DT) of hidden Markov models (HMMs) in speech recognition based on an efficient global optimization algorithm used to solve the so-called trust region (TR) problem, where a quadratic function is minimized under a spherical constraint. In the first method, maximum mutual information estimation (MMIE) of Gaussian mixture HMMs is formulated as a standard TR problem so that the efficient global optimization method can be used in each iteration to maximize the auxiliary function of discriminative training for speech recognition. In the second method, we propose to construct a new auxiliary function for DT of HMMs by adding a quadratic penalty term. The new auxiliary function is constructed to serve as first-order approximation as well as lower bound of the original discriminative objective function within a locality constraint. Due to the lower-bound property, the found optimal point of the new auxiliary function is guaranteed to improve the original discriminative objective function until it converges to a local optimum or stationary point of the objective function. Both TR-based optimization methods have been investigated on two standard large-vocabulary continuous speech recognition tasks, using the WSJ0 and Switchboard databases. Experimental results have shown that the proposed TR methods outperform the conventional EBW method in terms of convergence behavior as well as recognition performance.**

***Index Terms*—Discriminative training, global optimization, lower-bounded auxiliary function, trust region problem.**

## I. Introduction

RECENTLY, discriminative training (DT) methods have achieved tremendous success on a variety of automatic speech recognition (ASR) tasks. Many DT methods have been proposed to estimate Gaussian mixture hidden Markov models (HMMs), see a recent survey in [8]. Discriminative training of HMM parameters is a typical optimization problem. We first formulate an objective function according to certain discriminative criterion, such as maximum mutual information (MMI) [2], [22], minimum classification error (MCE) [9], minimum word or phone error (MWE or MPE) [16], large margin estimation (LME) [5], [10], [18], [21], etc. Next, an effective optimization method is used to minimize or maximize the objective function with respect to (w.r.t.) model parameters. In speech recognition, many different methods have been used to optimize the derived objective functions to estimate Gaussian mixture HMMs, including generalized probabilistic descent (GPD) [9] based on first-order gradient descent algorithm, approximate second-order Quickprop method [13], extended Baum–Welch (EBW) algorithm [2], [4], [22] based on growth transformation, constrained line search [11] and convex optimization [7]. The major difficulty of DT in ASR lies in the fact that the above-mentioned DT criteria normally lead to quite complicated objective functions, which are pretty hard to optimize directly. In many cases, we need to construct an auxiliary function in simpler form and then iteratively optimize the auxiliary function instead of the original objective function. From the viewpoint of optimization theory, two important issues must be addressed for this formulation: 1) how to effectively find the optimal point of the auxiliary function; 2) how to ensure the found optimal point of the auxiliary function also improves the original objective function. In the well-known EM algorithm [3], both problems are solved perfectly. In EM, the constructed auxiliary function is always a concave function so that its global maximum can be obtained efficiently by either convex optimization methods or even a closed-form solution in many cases. Moreover, the auxiliary function in EM is a strict lower-bound of the original objective function. As a result, any increase in the auxiliary function always corresponds to an even larger increase in the original objective function. Because of these nice properties, convergence of the objective function is guaranteed in EM. In other words, the original objective function is guaranteed to increase over EM iterations until it converges to a local optimum or stationary point of the original objective function. In discriminative training, an EM-style auxiliary function is normally constructed as in [6], [7], [17], but convergence of the corresponding iterative optimization methods is not theoretically guaranteed. The reason is twofold: 1) the derived auxiliary function in DT is neither convex nor concave so that optimization of the nonconvex auxiliary function itself is a big challenge; 2) the derived auxiliary function is not a strict lower-bound of the original DT objective function so that optimizing the auxiliary functions does not guarantee to improve the original objective function accordingly.

In this paper, we have proposed to use the so-called trust region (TR) method [14], [19] to address the above-mentioned

Manuscript received May 18, 2010; revised October 20, 2010; accepted April 12, 2011. Date of publication April 21, 2011; date of current version September 23, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mark J. F. Gales.

C. Liu and Y. Hu are with iFlytek Research, Hefei 230000, China (e-mail: ustc.congliu@gmail.com; congliu2@iflytek.com; yuhu@iflytek.com).

L.-R. Dai is with the iFlytek Speech Lab, University of Science and Technology of China, Hefei 230000, China (e-mail: lrdai@ustc.edu.cn).

H. Jiang is with the Department of Computer Science and Engineering, York University, Toronto, ON M3J 1P3, Canada (e-mail: hj@cse.yorku.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2144969

two issues in discriminative training of HMMs for speech recognition. Here, we refer TR problem as minimization of a quadratic function subject to a sphere or elliptic constraint[19]. The TR problem is one of a few non-convex optimization problems where the globally optimal solution can be found in an efficient way. In this work, we take MMI estimation as an example though the same ideas can be easily extended to other discriminative criteria. Following the approximation–maximization manner in [6], [7], we first derive an auxiliary function to approximate the MMI objective function of HMMs in a close neighborhood of initial model parameters [11]. For Gaussian mixture HMMs, the auxiliary function can be represented as an indefinite quadratic function of model parameters. Obviously, under the assumption of the locality constraint, the TR-based global optimization algorithm can be used to find the global optimum of the non-convex auxiliary function, which nicely solves the first critical issue arising in DT of HMMs as mentioned above. Furthermore, we propose to construct a new auxiliary function by penalizing the above auxiliary function with a positive definite quadratic term, which is derived by upper-bounding the gap function between the old auxiliary function and the MMI original objective function within the locality constraint. In this way, the new penalized auxiliary function serves as a strict lower bound of the MMI objective function and also takes the same function form as the old one so that it can still be optimized by the TR algorithm. More importantly, just like the EM algorithm, any increase in the new auxiliary function always guarantees an even larger increase of the original objective function due to the lower-bound property. Therefore, the second problem in DT as mentioned above has also been nicely solved. In this work, we have evaluated the proposed TR methods on two standard large-vocabulary ASR tasks using the WSJ0 and Switchboard databases. Experimental results have shown that the proposed TR methods yields much better learning curves than the conventional EBW method in terms of maximizing the MMI objective function in the MMI training process. It is also observed that the proposed TR methods can achieve better recognition performance than the conventional EBW method on all evaluated ASR tasks.

The remainder of this paper is organized as follows. In Section II, we briefly review the standard TR problem along with an efficient global optimization algorithm to solve TR problems. In Section III, we take the MMI training as an example to demonstrate how to formulate the MMI estimation of HMMs as a TR problem so that it can be efficiently solved by the efficient TR algorithm. In Section IV, we study how to use a quadratic penalty term to construct a new auxiliary function, which serves as a strict lower-bound of the MMI objective function. The new auxiliary function takes the same function form so that the efficient TR algorithm can be equally applied to find its globally optimal solution as well. Next, experimental results on two large vocabulary ASR tasks using the WSJ0 and Switchboard databases are reported and discussed in Section V. At last, the paper is concluded with our conclusions and findings.

## II. Trust Region Problem and Its Solution

It is well known that most non-convex optimization problems are difficult to solve. One of a few exceptions is minimization of an indefinite quadratic function under a spherical or elliptic constraint, which is usually called *trust region (TR)* problems [19]. In this section, we briefly review the optimization theory to show how the *global* optimum of the TR problem can be efficiently found using a fast global optimization algorithm.[1]

Recall that a general quadratic function w.r.t. a $n$-variable vector, $\mathbf{x}$, has the form $(1/2)\mathbf{x}^\top Q\mathbf{x} + \mathbf{q}^\top\mathbf{x}$, where $Q$ is a symmetric (not necessarily positive definite) matrix and $\mathbf{q}$ is a vector. A standard TR problem is expressed as

$$\begin{aligned} \min_{\mathbf{x}\in\mathbf{R}^n} \quad & \frac{1}{2}\mathbf{x}^\top Q\mathbf{x} + \mathbf{q}^\top\mathbf{x} \\ \text{s.t.} \quad & \mathbf{x}^\top\mathbf{x} \le \rho^2 \end{aligned} \tag{1}$$

with $\rho$ a constant to control size of the spherical trust region. If $Q$ is positive definite, the global minimum to (1) can be calculated as $\hat{\mathbf{x}} = -Q^{-1}\mathbf{q}$. Furthermore, if the norm of $\hat{\mathbf{x}}$ is bounded by $\rho^2$, i.e., $\hat{\mathbf{x}}^\top\hat{\mathbf{x}} \le \rho^2$, then $\hat{\mathbf{x}}$ is a feasible solution of the TR problem in (1). In all other situations, the global minimum of (1) can also be found efficiently based on the following theorem [14].

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

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

where $I$ denotes identity matrix.

As proven in [14], the conditions in (2) are both necessary and sufficient conditions of that $\mathbf{x}^*$ is the globally optimal solution to (1). Based on the first condition in (2), the global minimum $\mathbf{x}^*$ can be easily calculated based on a scalar $\tau$ as

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

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

Moreover, another theorem in [14] is useful for searching the optimal scalar $\tau$ for $\mathbf{x}^*$. If we define $\tau_0$ as the minimum value of $\tau$ such that $Q+\tau I$ is positive semi-definite, then it is easy to see that $\tau_0$ is equal to the negative value of the smallest (closest to -$\infty$) eigenvalue of $Q$.

*Theorem 2:* If $\mathbf{q} \neq 0$, $\tau_1$ and $\tau_2$ are two scalars that satisfy $\tau_0 \le \tau_1 < \tau_2$. Let $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ are solutions to $(Q+\tau_1 I)\mathbf{x}_1^* = -\mathbf{q}$ and $(Q+\tau_2 I)\mathbf{x}_2^* = -\mathbf{q}$, respectively, then $\|\mathbf{x}_1^*\|_2 > \|\mathbf{x}_2^*\|_2$.

In other words, the norm $\|(Q+\tau I)^{-1}\mathbf{q}\|_2$ is a monotonically decreasing function of $\tau$ for $\tau > \tau_0$. As a result, the unique scalar, denoted as $\tau^*$, which satisfies $\|(Q+\tau^* I)^{-1}\mathbf{q}\|_2 = \rho^2$, can be efficiently found in the interval $[\tau_0, +\infty)$ using a binary search method.

[1]Global optimization refers to optimization methods that are guaranteed to find a globally optimal solution.

## III. Formulating MMI Training as Trust Region Problem

### A. MMI Training as Constrained Optimization

Given a training set containing $R$ utterances $\{O_1, \ldots, O_R\}$ along with the corresponding transcripts $\{S_1, \ldots, S_R\}$, the well-known MMI objective function in speech recognition can be expressed as

$$\mathbf{F}_{\mathrm{MMI}}(\Lambda) = \sum_{r=1}^{R} \left[ \log p(O_r, S_r \mid \Lambda) - \log \sum_{s_r \in \mathcal{M}_r} p(O_r, s_r \mid \Lambda) \right] \tag{4}$$

where $\Lambda$ represents the set of all HMM parameters, and $\mathcal{M}_r$ denotes a word graph generated for $O_r$ consisting of all possible competing word sequences.

As shown in [11], it is beneficial to impose a locality constraint on model parameters $\Lambda$ during each iteration to ensure that they do not deviate too much from the initial value, i.e., $\Lambda^{(n)}$. The locality constraint can be quantitatively defined based on Kullback–Leibler (KL) divergence. Therefore, the MMI training of HMM parameters, $\Lambda$, can be formulated as the following iterative constrained maximization problem:

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

$$\text{subject to} \quad \mathcal{D}(\Lambda \parallel \Lambda^{(n)}) \le \rho^2 \tag{6}$$

where $\mathcal{D}(\Lambda \parallel \Lambda^{(n)})$ is the KL divergence between probability distributions specified by $\Lambda$ and $\Lambda^{(n)}$, and $\rho > 0$ is a constant to define the feasible trust region for optimization.

In the following, we consider how to convert the above constrained maximization problem in (5) and (6) into a standard TR problem as in (1).

### B. Reformulating KL-Based Constraint as Spherical Constraint

We will first show how to formulate the KL-based locality constraint in (6) as a spherical form, as required in the standard trust region problem in (1). Assume there are totally $K$ Gaussians in the HMM set, i.e., $\Lambda = \{\lambda_k | k = 1, \ldots, K\}$, where $\lambda_k$ denotes a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\Sigma_k$, i.e., $\mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$ with $k \in (1, \ldots, K)$. For simplicity, in this paper, we assume all covariance matrices, $\Sigma_k$, are diagonal. As shown in [11], the joint constraint in (6) can be relaxed as sum of individual constraints related to all Gaussians:

$$\mathcal{D}(\Lambda \parallel \Lambda^{(n)}) \le \sum_k \mathcal{D}(\lambda_k \parallel \lambda_k^{(n)}) \le \rho^2. \tag{7}$$

As we know, the KL divergence between two Gaussians, i.e., $\mathcal{D}(\lambda_k \parallel \lambda_k^{(n)})$, can be calculated by the following closed-form formula:

$$\begin{aligned} \mathcal{D}(\lambda_k \parallel \lambda_k^{(n)}) &= \mathcal{D}(\mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k) \parallel \mathcal{N}(\boldsymbol{\mu}_k^{(n)}, \Sigma_k^{(n)})) \\ &= \frac{1}{2}\Big[ (\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^{(n)})^\top \Sigma_k^{(n)-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^{(n)}) \\ &\quad + \mathrm{tr}\big(\Sigma_k \Sigma_k^{(n)-1}\big) + \log \frac{|\Sigma_k^{(n)}|}{|\Sigma_k|} - D \Big] \end{aligned} \tag{8}$$

where $D$ is dimension of feature vectors.

Furthermore, we can break down the constraint in (8) into two separate terms relevant to Gaussian mean and covariance matrix, respectively,

$$\mathcal{D}(\boldsymbol{\mu}_k \parallel \boldsymbol{\mu}_k^{(n)}) = \frac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^{(n)})^\top \Sigma_k^{(n)-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^{(n)}) \tag{9}$$

$$\mathcal{D}(\Sigma_k \parallel \Sigma_k^{(n)}) = \frac{1}{2}\Big[ \mathrm{tr}(\Sigma_k \Sigma_k^{(n)-1}) + \log \frac{|\Sigma_k^{(n)}|}{|\Sigma_k|} - D \Big]. \tag{10}$$

If we normalize each mean vector $\boldsymbol{\mu}_k$ with the initial mean vector, $\boldsymbol{\mu}_k^{(n)}$, and the initial covariance matrix, $\Sigma_k^{(n)}$, of the corresponding Gaussian as $\hat{\boldsymbol{\mu}}_k = \Sigma_k^{(n)-1/2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^{(n)})$, the locality constraint for Gaussian mean vectors in (9) can be easily formulated as a quadratic form as follows:

$$\sum_k \mathcal{D}(\boldsymbol{\mu}_k \parallel \boldsymbol{\mu}_k^{(n)}) \propto \mathbf{x}_{\boldsymbol{\mu}}^\top \mathbf{x}_{\boldsymbol{\mu}} \le \rho^2 \tag{11}$$

where $\mathbf{x}_{\boldsymbol{\mu}}$ denotes a large super-vector (in column) constructed with all normalized mean vectors as $\mathbf{x}_{\boldsymbol{\mu}} = \left[\hat{\boldsymbol{\mu}}_1^\top, \hat{\boldsymbol{\mu}}_2^\top, \ldots, \hat{\boldsymbol{\mu}}_K^\top\right]^\top_{(KD \times 1)}$.

Similarly, we consider to formulate the KL-based constraint for covariance matrix in (10) into the same spherical format. As mentioned above, we assume all covariance matrices $\Sigma_k$ are diagonal: $\Sigma_k = \mathrm{diag}(\sigma_{k1}^2, \sigma_{k2}^2, \ldots, \sigma_{kD}^2)$. Thus, the KL-based locality of covariance matrices in (10) can be simplified as

$$\mathcal{D}(\Sigma_k \parallel \Sigma_k^{(n)}) = \frac{1}{2} \sum_d \left[ \frac{\sigma_{kd}^2}{(\sigma_{kd}^{(n)})^2} - \log \frac{\sigma_{kd}^2}{(\sigma_{kd}^{(n)})^2} - 1 \right]. \tag{12}$$

If we normalize Gaussian variances using the following transformation: $\hat{\sigma}_{kd} = \log(\sigma_{kd}/\sigma_{kd}^{(n)})$ $(\forall k, d)$ based on the initial variance values, the constraint in (12) can be expressed as

$$\mathcal{D}(\Sigma_k \parallel \Sigma_k^{(n)}) = \frac{1}{2} \sum_d \left[ e^{2\hat{\sigma}_{kd}} - 2\hat{\sigma}_{kd} - 1 \right]. \tag{13}$$

Based on the second-order approximation of the Taylor series expansion in [11]: $e^y - y - 1\big|_{y \approx 0} \approx y^2/2$, the KL constraint for covariance matrix in (13) can be approximated as the following quadratic form:

$$\mathcal{D}(\Sigma_k \parallel \Sigma_k^{(n)}) \approx \sum_d \hat{\sigma}_{kd}^2. \tag{14}$$

Assume we define a variance vector using all normalized variances as $\hat{\boldsymbol{\sigma}}_k = [\hat{\sigma}_{k1}, \hat{\sigma}_{k2}, \ldots, \hat{\sigma}_{kD}]^\top$ for the $k$th Gaussian, and construct a super-vector $\mathbf{x}_\Sigma$ based on the normalized variance vectors from all Gaussians in the whole model set as $\mathbf{x}_\Sigma = \left[\hat{\boldsymbol{\sigma}}_1^\top, \hat{\boldsymbol{\sigma}}_2^\top, \ldots, \hat{\boldsymbol{\sigma}}_K^\top\right]^\top_{(KD \times 1)}$, we can derive a spherical constraint for covariance matrices as follows:

$$\sum_k \mathcal{D}(\Sigma_k \parallel \Sigma_k^{(n)}) \propto \mathbf{x}_\Sigma^\top \mathbf{x}_\Sigma \le \rho^2. \tag{15}$$

### C. Casting MMI Auxiliary Function as Quadratic Function

Following the iterative optimization strategy as the approximation–maximization (AM) method in [6] and [7], we construct an auxiliary function in each iteration and maximize the auxiliary function so that the original objective function can be optimized in an indirect way. As in [7], we first construct an auxiliary function $\mathcal{Q}$ as a local approximation of $\mathbf{F}_{\mathrm{MMI}}$ around an initial point $\Lambda^{(n)}$ as

$$\begin{aligned}
&\mathcal{Q}(\Lambda|\Lambda^{(n)}) \\
&= \sum_r \sum_{\mathbf{l}_r} p(\mathbf{l}_r \mid O_r, S_r, \Lambda^{(n)}) \cdot \log p(O_r, \mathbf{l}_r, S_r \mid \Lambda) \\
&\quad - \sum_r \sum_{s_r \in \mathcal{M}_r} \sum_{\mathbf{l}_r} p(\mathbf{l}_r \mid O_r, \mathcal{M}_r, \Lambda^{(n)}) \cdot \log p(O_r, \mathbf{l}_r, s_r \mid \Lambda) \\
&\quad + C_1 \\
&= \sum_k \sum_r \sum_t \left(\gamma_{krt}^{+} - \gamma_{krt}^{-}\right) \log p(\boldsymbol{o}_{rt}, k \mid \Lambda) + C_1
\end{aligned} \tag{16}$$

where $\mathbf{l}_r$ denotes the unobserved Gaussian label sequences of $O_r$ in HMM, and $\gamma_{krt}^{+}$ and $\gamma_{krt}^{-}$ denote occupancy statistics collected for $k$th Gaussian kernel based on reference $S_r$ and word graph $\mathcal{M}_r$, respectively, and $p(\boldsymbol{o}_{rt}, k \mid \Lambda)$ stands for output probability of one feature vector calculated based on $k$th Gaussian distribution in the model set, and $C_1$ is a constant independent of $\Lambda$.

It is straightforward to show that the above auxiliary function, $\mathcal{Q}(\Lambda|\Lambda^{(n)})$, satisfies $\mathbf{F}_{\mathrm{MMI}}(\Lambda)|_{\Lambda=\Lambda^{(n)}} = \mathcal{Q}(\Lambda|\Lambda^{(n)})|_{\Lambda=\Lambda^{(n)}}$, and $\partial \mathbf{F}_{\mathrm{MMI}}(\Lambda)/\partial\Lambda|_{\Lambda=\Lambda^{(n)}} = \partial \mathcal{Q}(\Lambda|\Lambda^{(n)})/\partial\Lambda|_{\Lambda=\Lambda^{(n)}}$. Therefore, $\mathcal{Q}(\Lambda|\Lambda^{(n)})$ can be viewed as a first-order tangential approximation of the MMI objective function, $\mathbf{F}_{\mathrm{MMI}}(\Lambda)$, at $\Lambda^{(n)}$.

Since we have

$$\log p(\boldsymbol{o}_{rt}, k|\Lambda) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\boldsymbol{o}_{rt} - \boldsymbol{\mu}_k)^{\top}\Sigma_k^{-1}(\boldsymbol{o}_{rt} - \boldsymbol{\mu}_k) + C_2$$

for Gaussian mixture HMMs, maximizing the auxiliary function $\mathcal{Q}(\Lambda|\Lambda^{(n)})$ in (16) is equivalent to minimizing the following function:

$$\begin{aligned}
\bar{\mathcal{Q}}(\Lambda|\Lambda^{(n)}) = \frac{1}{2} \sum_k \sum_r \sum_t \left(\gamma_{krt}^{+} - \gamma_{krt}^{-}\right) \\
\cdot \left[\log|\Sigma_k| + (\boldsymbol{o}_{rt} - \boldsymbol{\mu}_k)^{\top}\Sigma_k^{-1}(\boldsymbol{o}_{rt} - \boldsymbol{\mu}_k)\right].
\end{aligned} \tag{17}$$

### D. Update Gaussian Mean Vectors Using the TR Methods

If we assume covariance matrices are constant and only update Gaussian mean vectors, the objective function in (17) can be simplified as follows:

$$\begin{aligned}
&\bar{\mathcal{Q}}(\boldsymbol{\mu}|\boldsymbol{\mu}^{(n)}) \\
&= \frac{1}{2} \sum_k \sum_r \sum_t \left(\gamma_{krt}^{+} - \gamma_{krt}^{-}\right) \left[(\boldsymbol{o}_{rt} - \boldsymbol{\mu}_k)^{\top}\Sigma_k^{(n)-1}(\boldsymbol{o}_{rt} - \boldsymbol{\mu}_k)\right] \\
&= \frac{1}{2} \sum_k \sum_r \sum_t \left(\gamma_{krt}^{+} - \gamma_{krt}^{-}\right) \left[\hat{\boldsymbol{\mu}}_k^{\top}\hat{\boldsymbol{\mu}}_k - 2\hat{\boldsymbol{o}}_{krt}^{\top}\hat{\boldsymbol{\mu}}_k\right] + C_3
\end{aligned} \tag{18}$$

where $\hat{\boldsymbol{o}}_{krt} = \Sigma_k^{(n)-1/2}(\boldsymbol{o}_{rt} - \boldsymbol{\mu}_k^{(n)})$ denotes feature vectors normalized in the same manner using the initial Gaussian mean vector and covariance matrix.

Let $\gamma_k = \sum_r \sum_t (\gamma_{krt}^{+} - \gamma_{krt}^{-})$ and $\boldsymbol{\xi}_k = -\sum_r \sum_t (\gamma_{krt}^{+} - \gamma_{krt}^{-})\hat{\boldsymbol{o}}_{krt}$ denote the corresponding zero-order and first-order statistics collected from all available training data, we define the following matrix and vector as

$$Q_{\boldsymbol{\mu}} = \begin{bmatrix} \gamma_1 \cdot I_{D\times D} & & & \\ & \gamma_2 \cdot I_{D\times D} & & \\ & & \ddots & \\ & & & \gamma_K \cdot I_{D\times D} \end{bmatrix}_{(KD\times KD)}$$

$$\mathbf{q}_{\boldsymbol{\mu}} = \left[\boldsymbol{\xi}_1^{\top}, \boldsymbol{\xi}_2^{\top}, \ldots, \boldsymbol{\xi}_K^{\top}\right]_{(KD\times 1)}^{\top}.$$

MMI estimation of Gaussian mean vectors turns into a constrained minimization problem as follows:

$$\begin{aligned}
\min_{\boldsymbol{\mu}} \ \bar{\mathcal{Q}}(\boldsymbol{\mu}|\boldsymbol{\mu}^{(n)}) &= \min_{\mathbf{x}_{\boldsymbol{\mu}}} \left[\frac{1}{2}\mathbf{x}_{\boldsymbol{\mu}}^{\top} Q_{\boldsymbol{\mu}} \mathbf{x}_{\boldsymbol{\mu}} + \mathbf{q}_{\boldsymbol{\mu}}^{\top}\mathbf{x}_{\boldsymbol{\mu}}\right] \\
\text{s.t.} \quad \mathbf{x}_{\boldsymbol{\mu}}^{\top}\mathbf{x}_{\boldsymbol{\mu}} &\le \rho^2.
\end{aligned} \tag{19}$$

where $\mathbf{x}_{\boldsymbol{\mu}}$ is the super-vector of all normalized mean vectors as constructed in (11). Obviously, this is a standard trust region problem as defined in (1).

### E. Update Covariance Matrices Using the TR Methods

On the other hand, if we assume Gaussian mean vectors are constant and only update the diagonal covariance matrices, (17) can be simplified as follows:

$$\begin{aligned}
\bar{\mathcal{Q}}(\Sigma|\Sigma^{(n)}) &= \frac{1}{2} \sum_k \sum_r \sum_t \left(\gamma_{krt}^{+} - \gamma_{krt}^{-}\right) \\
&\quad \cdot \left[\log|\Sigma_k| + (\boldsymbol{o}_{rt} - \boldsymbol{\mu}_k^{(n)})^{\top}\Sigma_k^{-1}(\boldsymbol{o}_{rt} - \boldsymbol{\mu}_k^{(n)})\right] \\
&= \frac{1}{2} \sum_k \sum_r \sum_t \left(\gamma_{krt}^{+} - \gamma_{krt}^{-}\right) \\
&\quad \cdot \sum_d \left[2 \cdot \log\frac{\sigma_{kd}}{\sigma_{kd}^{(n)}} + \left(\frac{\sigma_{kd}^{(n)}}{\sigma_{kd}}\right)^2 \cdot \hat{o}_{krtd}^2\right] + C_4 \\
&\approx \frac{1}{2} \sum_k \sum_r \sum_t \left(\gamma_{krt}^{+} - \gamma_{krt}^{-}\right) \\
&\quad \cdot \sum_d \left[2\hat{\sigma}_{kd} + (1 - 2\hat{\sigma}_{kd} + 2\hat{\sigma}_{kd}^2) \cdot \hat{o}_{krtd}^2\right] + C_4 \\
&= \sum_k \sum_r \sum_t \left(\gamma_{krt}^{+} - \gamma_{krt}^{-}\right) \\
&\quad \cdot \sum_d \left[\hat{o}_{krtd}^2 \cdot \hat{\sigma}_{kd}^2 + (1 - \hat{o}_{krtd}^2)\hat{\sigma}_{kd}\right] + C_5
\end{aligned} \tag{20}$$

where $\hat{o}_{krtd}$ denotes the $d$th dimension of the normalized feature vector $\hat{\boldsymbol{o}}_{krt}$.

If we denote the corresponding statistics as $\eta_{kd} = 2\sum_r \sum_t (\gamma_{krt}^{+} - \gamma_{krt}^{-}) \cdot \hat{o}_{krtd}^2$, $\zeta_{kd} = \sum_r \sum_t (\gamma_{krt}^{+} - \gamma_{krt}^{-})(1 -$

$\hat{o}^2_{krtd}$), we can define the following matrix and vector in a similar way as follows:

$$Q_\Sigma = \begin{bmatrix} \eta_{11} & & & & & & \\ & \eta_{12} & & & & & \\ & & \ddots & & & & \\ & & & \eta_{1D} & & & \\ & & & & \eta_{21} & & \\ & & & & & \ddots & \\ & & & & & & \eta_{KD} \end{bmatrix}_{(KD\times KD)}$$

$$\mathbf{q}_\Sigma = [\zeta_{11}, \zeta_{12}, \ldots, \zeta_{KD}]^\top_{(KD\times 1)}.$$

Therefore, the MMI estimation of Gaussian covariance matrices in each iteration can be converted into the following TR problem as

$$\min_{\Sigma} \bar{\mathcal{Q}}(\Sigma|\Sigma^{(n)}) = \min_{\mathbf{x}_\Sigma} \left[\frac{1}{2}\mathbf{x}_\Sigma^\top Q_\Sigma \mathbf{x}_\Sigma + \mathbf{q}_\Sigma^\top \mathbf{x}_\Sigma\right] \quad \text{s.t.} \quad \mathbf{x}_\Sigma^\top \mathbf{x}_\Sigma \le \rho^2. \tag{21}$$

In summary, the MMI objective function can be approximated by the above auxiliary function $\mathcal{Q}(\Lambda|\Lambda^{(n)})$ in (16). Therefore, the MMI estimation of either Gaussian mean vectors or covariance matrices of HMMs can be converted into an iterative optimization problem, which is a standard TR problem with indefinite matrices $Q_{\boldsymbol{\mu}}$ and $Q_\Sigma$. Therefore, the globally optimal solution in each iteration can be easily derived as

$$\mathbf{x}^*_{\boldsymbol{\mu}} = -(Q_{\boldsymbol{\mu}} + \tau^*_{\boldsymbol{\mu}} I)^{-1}\mathbf{q}_{\boldsymbol{\mu}} \tag{22}$$

$$\mathbf{x}^*_\Sigma = -(Q_\Sigma + \tau^*_\Sigma I)^{-1}\mathbf{q}_\Sigma \tag{23}$$

where the optimal values of $\tau^*_{\boldsymbol{\mu}}$ and $\tau^*_\Sigma$ can be efficiently found by the binary search algorithm as introduced in Section II.

As a remark, the ideas in this section have recently been studied in [12] as well. The major difference is that we use a more efficient binary search for the optimal value of $\tau$, instead of the second order Newton method as in [12].

## IV. MMI as Bounded Trust Region Problem

Under the framework of locality constraint, the MMI objective function is first approximated by a simpler auxiliary function, which is optimized in each iteration instead of the original MMI objective function. Even though the auxiliary function is neither convex nor concave in discriminative training, the proposed TR method has perfectly solved its optimization problem since the TR algorithm can efficiently find the globally optimal point of the nonconvex auxiliary function. However, the second problem in DT remains open: the found global optimum is not guaranteed to improve the original MMI objective function all the time since the tangential auxiliary function in (16) is not a strict lower bound of the MMI objective function.

### A. Lower-Bounded Auxiliary Function

In this study, we propose to construct a new auxiliary function by penalizing the auxiliary function in (16), $\mathcal{Q}(\Lambda|\Lambda^{(n)})$, by subtracting a positive definite quadratic term. The quadratic penalty term is constructed in such a way that the new auxiliary function remains as the first-order approximation of the original objective function and more importantly it also serves as a lower bound of the original objective function within the locality constraint. Furthermore, the new auxiliary function takes the exactly same function form as the old auxiliary function, $\mathcal{Q}(\Lambda|\Lambda^{(n)})$. Therefore, the TR-based global optimization algorithm can be similarly applied to find the globally optimal point of the new auxiliary function. The critical difference is that the found global solution is guaranteed to improve the original MMI objective function because of the lower-bound property.

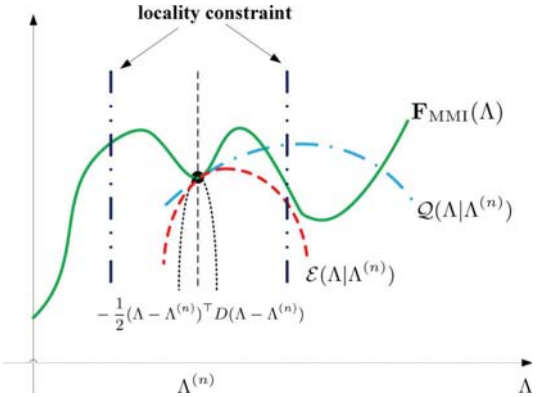


Fig. 1. Illustration of the lower bounded auxiliary function in discriminative training.

As illustrated in Fig. 1, we propose to construct a new auxiliary function by penalizing $\mathcal{Q}(\Lambda|\Lambda^{(n)})$ as follows:

$$\mathcal{E}(\Lambda|\Lambda^{(n)}) = \mathcal{Q}(\Lambda|\Lambda^{(n)}) - \frac{1}{2}(\Lambda-\Lambda^{(n)})^\top P(\Lambda-\Lambda^{(n)}) \tag{24}$$

where $P$ is a positive semi-definite matrix, denoted as $P \succeq 0$. Obviously, the new auxiliary function $\mathcal{E}(\Lambda|\Lambda^{(n)})$ still remains as a tangential first-order approximation of the original MMI objective function since it is straightforward to prove that $\mathbf{F}_{\mathrm{MMI}}(\Lambda)\Big|_{\Lambda=\Lambda^{(n)}} = \mathcal{E}(\Lambda|\Lambda^{(n)})\Big|_{\Lambda=\Lambda^{(n)}}$ and $\partial\mathbf{F}_{\mathrm{MMI}}(\Lambda)/\partial\Lambda\Big|_{\Lambda=\Lambda^{(n)}} = \partial\mathcal{E}(\Lambda|\Lambda^{(n)})/\partial\Lambda\Big|_{\Lambda=\Lambda^{(n)}}$ based on the tangent property of $\mathcal{Q}(\Lambda|\Lambda^{(n)})$.

Moreover, because matrix $P$ is positive semi-definite, if $|P|$ is sufficiently large, as shown in Fig. 1, the added penalty term will drag down $\mathcal{Q}$ to make $\mathcal{E}$ a strict lower bound of the original objective function as

$$\mathcal{E}(\Lambda|\Lambda^{(n)}) \le \mathbf{F}_{\mathrm{MMI}}(\Lambda). \tag{25}$$

Since the penalty term is also quadratic form, $\mathcal{E}(\Lambda|\Lambda^{(n)})$ has the exactly same function form as $\mathcal{Q}(\Lambda|\Lambda^{(n)})$. If we represent $\mathcal{E}(\Lambda|\Lambda^{(n)})$ in terms of normalized super-vector $\mathbf{x}$ as above, optimization of the new auxiliary function in each iteration can be represented as the following TR problem:

$$\Lambda^* = \arg\max_{\Lambda} \mathcal{E}(\Lambda|\Lambda^{(n)}) = \arg\min_{\mathbf{x}} \left[\frac{1}{2}\mathbf{x}^\top(Q+P)\mathbf{x} + \mathbf{q}^\top\mathbf{x}\right] \tag{26}$$

subject to $\mathbf{x}^\top\mathbf{x} \le \rho^2$. Therefore, the above TR algorithm can be applied to solve this problem in the same way except that matrix $Q$ is replaced by $Q+P$.

### B. Derivation of Penalty Matrix $P$

Even though a sufficiently positive $P$ matrix can always compensate $\mathcal{Q}$ to make $\mathcal{E}$ a lower bound of $\mathbf{F}$, it is not desirable to add a too positive $P$ since it will result in a very slow convergence rate for model training. In this work, we propose to add a minimum $P$ that is just positive enough to ensure that (25) holds within the local region as specified in (6). In this section, we consider how to derive such a tight minimum positive definite matrix $P$ by upper-bounding a gap function between $\mathcal{Q}$ and $\mathbf{F}$.

As shown in the Appendix, the gap function between $\mathcal{G}(\Lambda|\Lambda^{(n)})$ between $\mathcal{Q}$ and $\mathbf{F}$ can be computed as

$$\begin{aligned}&\mathcal{G}(\Lambda|\Lambda^{(n)})\\&=\mathcal{Q}(\Lambda|\Lambda^{(n)})-\mathbf{F}_{\mathrm{MMI}}(\Lambda|\Lambda^{(n)})\\&=\sum_r\sum_{\mathbf{l}_r}p(\mathbf{l}_r|O_r,S_r,\Lambda^{(n)})\log\frac{p(\mathbf{l}_r|O_r,S_r,\Lambda)}{p(\mathbf{l}_r|O_r,S_r,\Lambda^{(n)})}\\&\quad-\sum_r\sum_{\mathbf{l}_r\in\mathcal{M}_r}p(\mathbf{l}_r|O_r,\mathcal{M}_r,\Lambda^{(n)})\log\frac{p(\mathbf{l}_r|O_r,\mathcal{M}_r,\Lambda)}{p(\mathbf{l}_r|O_r,\mathcal{M}_r,\Lambda^{(n)})}.\end{aligned}\tag{27}$$

We first apply the Bayes' theorem, i.e., $p(\mathbf{l}_r|O_r,S_r,\Lambda)=p(\mathbf{l}_r,O_r\mid S_r,\Lambda)/p(O_r\mid S_r,\Lambda)$, to the posterior probability terms in (27). Meanwhile, we make the two assumptions regarding model parameters during optimization as $p(O_r\mid S_r,\Lambda)\approx p(O_r\mid S_r,\Lambda^{(n)})$ and $\sum_{s_r\in\mathcal{M}_r}p(O_r\mid s_r,\Lambda)\approx\sum_{s_r\in\mathcal{M}_r}p(O_r\mid s_r,\Lambda^{(n)})$. Obviously, these two assumptions can be easily justified in the above iterative constrained optimization framework. Because of the locality constraint in (6), it is valid to assume that model parameters do not dramatically change from their initial value $\Lambda^{(n)}$ in each iteration. Therefore, the gap function $\mathcal{G}(\Lambda|\Lambda^{(n)})$ can be simplified as follows:

$$\begin{aligned}\mathcal{G}(\Lambda|\Lambda^{(n)})&=\sum_r\sum_{\mathbf{l}_r}p(\mathbf{l}_r|O_r,S_r,\Lambda^{(n)})\\&\quad\cdot\Big[\log p(\mathbf{l}_r,O_r|S_r,\Lambda)-\log p(\mathbf{l}_r,O_r|S_r,\Lambda^{(n)})\Big]\\&\quad-\sum_r\sum_{s_r\in\mathcal{M}_r}\sum_{\mathbf{l}_r}p(\mathbf{l}_r|O_r,\mathcal{M}_r,\Lambda^{(n)})\\&\quad\cdot\Big[\log p(\mathbf{l}_r,O_r|s_r,\Lambda)-\log p(\mathbf{l}_r,O_r|s_r,\Lambda^{(n)})\Big]\\&=\sum_k\sum_r\sum_t\left(\gamma_{krt}^{+}-\gamma_{krt}^{-}\right)\\&\quad\cdot\Big[\log p(\boldsymbol{o}_{rt},k\mid\Lambda)-\log p(\boldsymbol{o}_{rt},k\mid\Lambda^{(n)})\Big]\end{aligned}\tag{28}$$

where $\gamma_{krt}^{+}$ and $\gamma_{krt}^{-}$ denote statistics collected in the same way as in (16). After we substitute Gaussian distribution into (28), we can represent the gap function w.r.t. Gaussian means and covariances as shown in (29) at the bottom of the page.

In the following, we will derive matrix $P$ for the lower-bounded auxiliary functions for Gaussian mean vectors and covariance matrices, respectively.

*1) Deriving Matrix $P_{\boldsymbol{\mu}}$ for Gaussian Mean Vectors:* If we only update Gaussian mean vectors, we can assume covariance matrices remain constant, i.e., $\Sigma_k=\Sigma_k^{(n)}\ (\forall k)$, the gap function in (29) can be represented as a function of mean vectors:

$$\begin{aligned}\mathcal{G}(\boldsymbol{\mu}|\boldsymbol{\mu}^{(n)})&=-\frac{1}{2}\sum_k\sum_r\sum_t(\gamma_{krt}^{+}-\gamma_{krt}^{-})\\&\quad\cdot\left[(\hat{\boldsymbol{o}}_{krt}-\hat{\boldsymbol{\mu}}_k)^{\top}(\hat{\boldsymbol{o}}_{krt}-\hat{\boldsymbol{\mu}}_k)-\hat{\boldsymbol{o}}_{krt}^{\top}\hat{\boldsymbol{o}}_{krt}\right]\\&=-\frac{1}{2}\sum_k\sum_r\sum_t(\gamma_{krt}^{+}-\gamma_{krt}^{-})\\&\quad\cdot\left[\hat{\boldsymbol{\mu}}_k^{\top}\hat{\boldsymbol{\mu}}_k-2\hat{\boldsymbol{o}}_{krt}^{\top}\hat{\boldsymbol{\mu}}_k\right]\end{aligned}\tag{30}$$

where $\hat{\boldsymbol{\mu}}_k$ and $\hat{\boldsymbol{o}}_{krt}$ denote normalized mean vectors and feature vectors as defined above.

Since we only update mean vectors, the penalized term in (24) can also be represented as a function of mean vectors, i.e., $(1/2)\mathbf{x}_{\boldsymbol{\mu}}^{\top}P_{\boldsymbol{\mu}}\mathbf{x}_{\boldsymbol{\mu}}$. The key idea to derive matrix $P_{\boldsymbol{\mu}}$ is to ensure that the penalty term remains as a strict upper bound of the above gap function within the locality constraint $\mathbf{x}_{\boldsymbol{\mu}}^{\top}\mathbf{x}_{\boldsymbol{\mu}}\le\rho^2$ as

$$\frac{1}{2}\mathbf{x}_{\boldsymbol{\mu}}^{\top}P_{\boldsymbol{\mu}}\mathbf{x}_{\boldsymbol{\mu}}\ge\mathcal{G}(\boldsymbol{\mu}|\boldsymbol{\mu}^{(n)}).\tag{31}$$

It is easy to verify that (31) implies (25).

In order to derive simple closed-form solution for matrix $P$, we assume matrix $P$ is diagonal and use $p_{kd}$ to represent the corresponding diagonal element of matrix $P$ for the $d$th dimension of the $k$th Gaussian. Furthermore, we decompose the joint locality constraint $\mathbf{x}_{\boldsymbol{\mu}}^{\top}\mathbf{x}_{\boldsymbol{\mu}}\le\rho^2$ to all individual Gaussian dimensions as

$$\hat{\mu}_{kd}^2\le\rho_{kd}^2\ \ (\forall k,d)\tag{32}$$

where $\sum_k\sum_d\rho_{kd}^2=\rho^2$. Obviously, the decomposed constraints are stronger and tighter than the original joint constraint. Based on these, we will be able to derive a simple closed-form formula to calculate each diagonal element $p_{kd}$ of matrix $P$ independently. Note that in discriminative training of HMMs, the quadratic form in the TR problem, i.e., matrix $Q$, is usually indefinite. Therefore, the optimal solution normally appears on the outer surface of the locality constraint. A sufficient condition to calculate the minimum value for $p_{kd}$ is to ensure that the every Gaussian dimension in both sides of the condition in (31) satisfies on the outer surface of (32) as follows:

$$p_{kd}\hat{\mu}_{kd}+\gamma_k\hat{\mu}_{kd}+2\xi_{kd}\ge0\tag{33}$$

$$\hat{\mu}_{kd}^2=\rho_{kd}^2\tag{34}$$

where $\gamma_k=\sum_r\sum_t(\gamma_{krt}^{+}-\gamma_{krt}^{-})$, $\xi_{kd}=-\sum_r\sum_t(\gamma_{krt}^{+}-\gamma_{krt}^{-})\cdot\hat{o}_{krtd}$, and $\rho_{kd}$ is derived by allocating the global con-

$$\mathcal{G}(\Lambda|\Lambda^{(n)})=-\frac{1}{2}\sum_k\sum_r\sum_t(\gamma_{krt}^{+}-\gamma_{krt}^{-})\cdot\left[\log\frac{|\Sigma_k|}{|\Sigma_k^{(n)}|}+(\boldsymbol{o}_{rt}-\boldsymbol{\mu}_k)^{\top}\Sigma_k^{-1}(\boldsymbol{o}_{rt}-\boldsymbol{\mu}_k)-(\boldsymbol{o}_{rt}-\boldsymbol{\mu}_k^{(n)})^{\top}\Sigma_k^{(n)-1}(\boldsymbol{o}_{rt}-\boldsymbol{\mu}_k^{(n)})\right]\tag{29}$$

straint $\rho^2$ to all individual Gaussians in proportional to their statistics: $\min\{\gamma_k^+, \gamma_k^-, |\gamma_k^+ - \gamma_k^-|\}$ and subject to the condition $\sum_k \sum_d \rho_{kd}^2 = \rho^2$, and then distributing it to all Gaussian dimensions uniformly.

After substituting (34) into (33), the minimum value of $p_{kd}$ can be derived as follows:

$$p_{kd} = \max\left\{-\gamma_k + \frac{2\xi_{kd}}{\rho_{kd}}, -\gamma_k - \frac{2\xi_{kd}}{\rho_{kd}}, \epsilon\right\} \tag{35}$$

where $\epsilon$ ($\epsilon > 0$) is a small positive value to ensure matrix $P_{\boldsymbol{\mu}}$ positive definite. Finally, matrix $P_{\boldsymbol{\mu}}$ is constructed for Gaussian means as $P_{\boldsymbol{\mu}} = \alpha \cdot \mathrm{diag}\{p_{11}, \ldots, p_{1d}, \ldots, p_{kd}, \ldots p_{KD}\}$, where $\alpha$ ($\alpha > 0$) is a control parameter introduced to compensate all approximation and assumptions in deriving matrix $P$ as a lower bound of the original objective function.

*2) Deriving Matrix $P_\Sigma$ for Covariance Matrices:* If we only update covariance matrices in discriminative training, we assume Gaussian mean vectors are constant, i.e., $\boldsymbol{\mu}_k = \boldsymbol{\mu}_k^{(n)}$ $(\forall k)$, the gap function in (29) can be represented as a function of covariance matrices. Furthermore, if we still accept the second-order Taylor series approximation, we can simplify the gap function as follows:

$$\begin{aligned}
&\mathcal{G}(\Sigma|\Sigma^{(n)}) \\
&= -\frac{1}{2}\sum_k\sum_r\sum_t(\gamma_{krt}^+ - \gamma_{krt}^-) \\
&\quad \cdot \left[\log\frac{|\Sigma_k|}{|\Sigma_k^{(n)}|} + (\boldsymbol{o}_{rt} - \boldsymbol{\mu}_k^{(n)})^\top(\Sigma_k^{-1} - \Sigma_k^{(n)^{-1}})(\boldsymbol{o}_{rt} - \boldsymbol{\mu}_k^{(n)})\right] \\
&= -\frac{1}{2}\sum_k\sum_r\sum_t(\gamma_{krt}^+ - \gamma_{krt}^-) \\
&\quad \cdot \sum_d\left[2 \cdot \log\frac{\sigma_{kd}}{\sigma_{kd}^{(n)}} + \hat{o}_{krtd}^2 \cdot \left[\left(\frac{\sigma_{kd}^{(n)}}{\sigma_{kd}}\right)^2 - 1\right]\right] \\
&= -\frac{1}{2}\sum_k\sum_r\sum_t(\gamma_{krt}^+ - \gamma_{krt}^-) \\
&\quad \cdot \sum_d\left[2\hat{\sigma}_{kd} + \hat{o}_{krtd}^2 \cdot (e^{-2\hat{\sigma}_{kd}} - 1)\right] \\
&\approx -\frac{1}{2}\sum_k\sum_r\sum_t(\gamma_{krt}^+ - \gamma_{krt}^-) \\
&\quad \cdot \sum_d\left[2\hat{\sigma}_{kd} + \hat{o}_{krtd}^2 \cdot (2\hat{\sigma}_{kd}^2 - 2\hat{\sigma}_{kd})\right] \\
&= -\sum_k\sum_r\sum_t(\gamma_{krt}^+ - \gamma_{krt}^-) \\
&\quad \cdot \sum_d\left[\hat{o}_{krtd}^2 \cdot \hat{\sigma}_{kd}^2 + (1 - \hat{o}_{krtd}^2)\hat{\sigma}_{kd}\right].
\end{aligned} \tag{36}$$

Since we only consider to update covariance matrices, the penalized term in (24) can also be represented as a function of covariance matrices in matrix format as $(1/2)\mathbf{x}_\Sigma^\top P_\Sigma \mathbf{x}_\Sigma$, where $\mathbf{x}_\Sigma$ denotes the super-vector constructed by normalized variance vectors as in (15). The key idea to derive matrix $P_\Sigma$ is to derive the penalty term as an upper-bound of the gap function within the locality constraint $\mathbf{x}_\Sigma^\top \mathbf{x}_\Sigma \le \rho_\Sigma^2$ as

$$\frac{1}{2}\mathbf{x}_\Sigma^\top P_\Sigma \mathbf{x}_\Sigma \ge \mathcal{G}(\Sigma|\Sigma^{(n)}). \tag{37}$$

Following the same ideas to decompose the constraint to all Gaussian dimensions, we can compute the minimum value of $p_{kd}$ of all diagonal elements in matrix $P_\Sigma$, which satisfies (37), by solving the following two equations:

$$p_{kd}\hat{\sigma}_{kd} + \eta_{kd}\hat{\sigma}_{kd} + 2\zeta_{kd} \ge 0 \tag{38}$$

$$\hat{\sigma}_{kd}^2 = \rho_{kd}^2 \tag{39}$$

where $\eta_{kd} = 2\sum_r\sum_t(\gamma_{krt}^+ - \gamma_{krt}^-) \cdot \hat{o}_{krtd}^2$, $\zeta_{kd} = \sum_r\sum_t(\gamma_{krt}^+ - \gamma_{krt}^-)(1 - \hat{o}_{krtd}^2)$, and $\rho_{kd}$ is allocated from the global constraint $\rho$ in the same manner as the case of Gaussian mean vectors.

After substituting (39) to (38), the minimum value of $p_{kd}$ for each Gaussian dimension in matrix $P$ can be derived as follows:

$$p_{kd} = \max\left\{-\eta_{kd} + \frac{2\zeta_{kd}}{\rho_{kd}}, -\eta_{kd} - \frac{2\zeta_{kd}}{\rho_{kd}}, \epsilon\right\} \tag{40}$$

where $\epsilon$ ($\epsilon > 0$) is a small positive number to ensure positive definiteness of matrix $P_\Sigma$. At last, matrix $P_\Sigma$ is constructed for Gaussian covariance matrices as $P_\Sigma = \alpha \cdot \mathrm{diag}\{p_{11}, \ldots, p_{1d}, \ldots, p_{kd}, \ldots p_{KD}\}$, where $\alpha$ is the control parameter as above.

After deriving matrices $P_{\boldsymbol{\mu}}$ and $P_\Sigma$ for Gaussian means and covariance matrices, respectively, we can substitute them into (26) to construct the new auxiliary function for Gaussian mean vectors and covariance matrices, which can be optimized by using the same TR algorithm as introduced in Section II. In summary, for any pre-defined trust region $\rho$, we can calculate the corresponding penalized matrix $P$, which ensures that the new auxiliary function serves as a strict lower bound of the original objective function within the trust region specified by $\rho$. As a result, the proposed bounded TR method will converge to a local optimum or stationary point of the original objective function.

## V. Experiments

In this section, we evaluate effectiveness of the proposed TR-based optimization methods in discriminative training of HMMs on two large-vocabulary continuous speech recognition tasks using the Wall Street Journal (WSJ0) and Switchboard databases. In the experiments, we compare the proposed TR methods, namely the original TR algorithm as well as the bounded TR using the new auxiliary function, with the conventional EBW based optimization method. Experimental setup is summarized in Table I. In our discriminative training methods, we always use the best MLE models as initial models, which are estimated using the HTK toolkit based on a standard training procedure. In the EBW method, as suggested in [17], i-smoothing is applied and all parameters are fine tuned and their optimal values are set as $E = 2$, $\tau = 100$ (i-smoothing factor).

In the TR-based optimization methods, we need to tune parameter $\rho$ that controls the range of trust region for updating model parameters. In the proposed bounded TR method, we also need to tune the compensation parameter $\alpha$ in constructing matrix $P$.

In the following, we first conduct a few sets of experiments on the WSJ-5k task, including: 1) evaluating effect of different $\rho$ values in TR-based optimization methods; 2) tuning scaling

TABLE I
EXPERIMENTAL SETUP IS LISTED FOR ALL RECOGNITION TASKS

| Training set | | Acoustic features | training data | # tied states | mix/state |
|---|---|---|---|---|---|
| WSJ0 | | 13 MFCCs $+\Delta+\Delta\Delta$ | 15 hrs | 2774 | 8 |
| Switch- | 60-hour subset | 13 PLPs $+\Delta+\Delta\Delta$ | 60 hrs | 6000 | 8 |
| board | *h5train00* | 13 PLPs $+\Delta+\Delta\Delta$ | 265 hrs | 9000 | 40 |

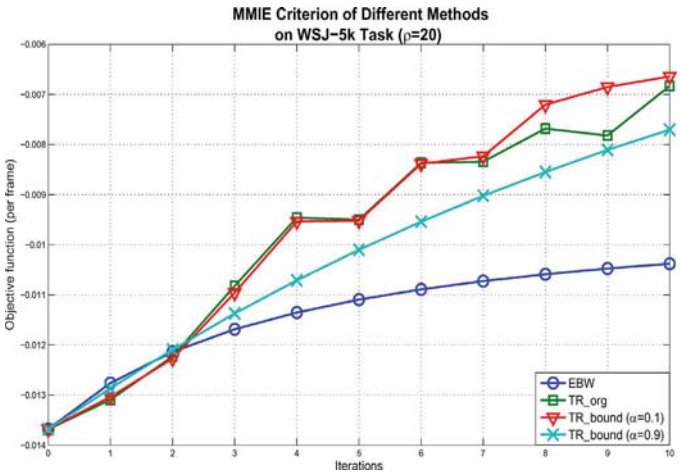


Fig. 2. Comparison of MMI objective function for different optimization methods on WSJ-5k task ($\rho = 20$).

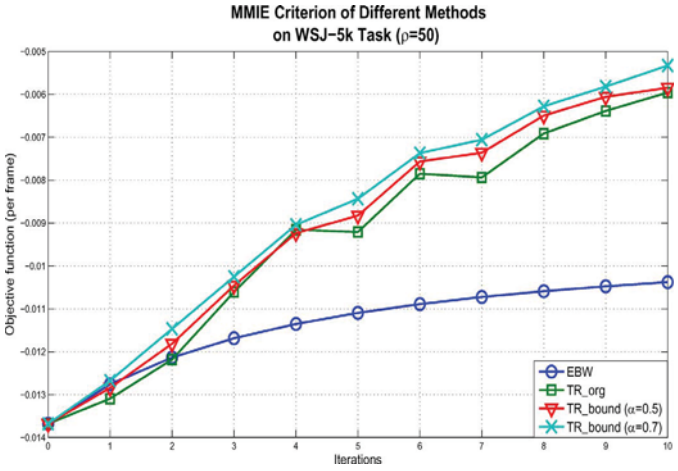


Fig. 3. Comparison of MMI objective function for different optimization methods on WSJ-5k task ($\rho = 50$).

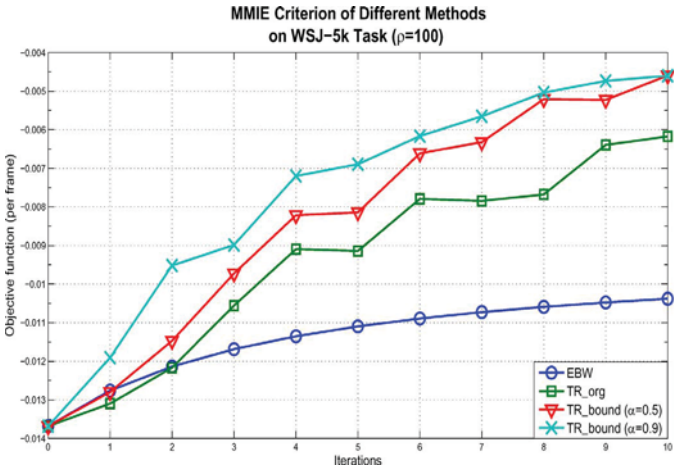


Fig. 4. Comparison of MMI objective function for different optimization methods on WSJ-5k task ($\rho = 100$).

factor $\alpha$ for the best convergency behavior; 3) comparing learning curves and recognition performance for different optimization methods when updating Gaussian means only or update both Gaussian means and covariance matrices at the same time. Note that in the EBW, we can update both mean and covariance matrices simultaneously in a single iteration. However, for the TR methods, we have to conduct two sub-steps: the first one is to update means only and the second one is to update covariance matrices only based on the re-collected statistics. Based on the best experimental setting in the WSJ-5k task, we further examine the TR methods in two more challenging tasks using the Switchboard database, namely the SWB 60-hour subset and the SWD *h5train00* full set.

### A. *WSJ-5k Task*

In this section, we evaluate the proposed TR methods on a small-scale task using the WSJ0 database. The training set is the standard SI-84 set, consisting of 7133 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. Note that this test set is used as both development set and evaluation set in our WSJ-5k experiments. For the MLE baseline, we use HTK to build cross-word triphone HMMs with a total number of 2,774 tied-states and 8 Gaussians per state. The word error rate (WER) of the MLE baseline using a standard trigram LM is 4.89%. The acoustic scaling factor used for DT is set as $\kappa = 1/15$.

We first tune parameters $\rho$ and $\alpha$ based on the configuration of updating Gaussian means only. We examine three different $\rho$ values for trust region: 20, 50 and 100. For each $\rho$ value, we also study the effect of different $\alpha$ values in the range [0.0, 2.0]. The learning curves of the MMI objective function using various optimization methods under different experimental setting are shown in Figs. 2–4, respectively. From the learning curves shown in Figs. 2–4, it is clear that both TR methods can significantly outperform the conventional EBW method in terms of optimizing the MMI objective function, under different choices of $\rho$. Furthermore, when setting a proper value for $\alpha$, the bounded TR method yields a much smoother and steadily growing learning curve comparing with that of the original TR method without the penalization term. From these results, we can see that a smaller scaling factor $\alpha$ is enough to achieve stable learning curves when the range of trust region $\rho$ is small, e.g.,

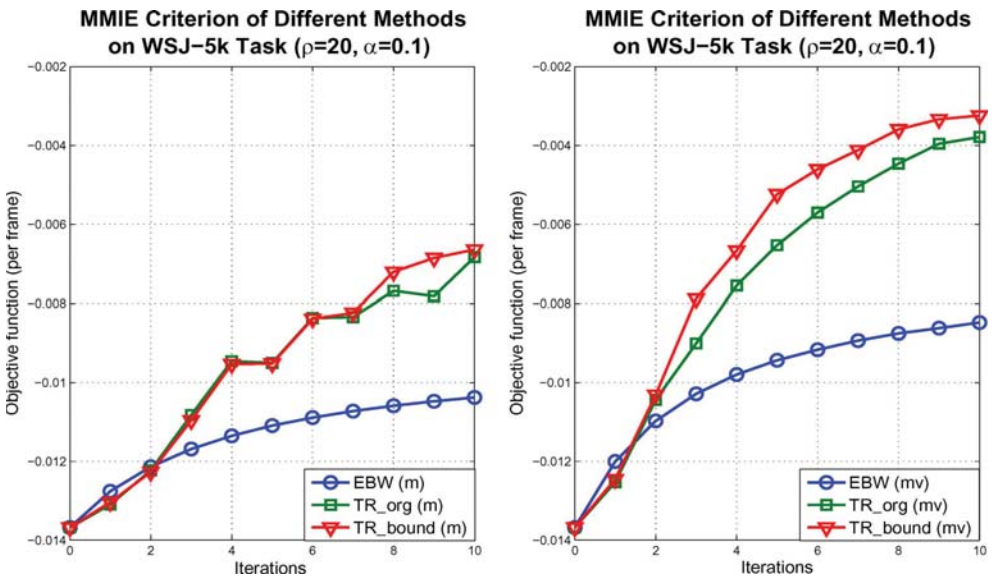


Fig. 5. Comparison of MMI objective function for different optimization methods on WSJ-5k: Updating Gaussian means only (denoted as $m$ in left figure) versus updating both Gaussian means and variances (denoted as $mv$ in right figure).

TABLE II
SUMMARY OF BEST RECOGNITION PERFORMANCE (WER IN %) ON SEVERAL ASR TASKS

| WER (%) | Test Set | Configuration | MLE | EBW | TR_org | TR_bound |
|---|---|---|---|---|---|---|
| WSJ-5k | Nov92 | m | 4.89 | 4.30 | 3.75 | 3.72 |
| | | m & v | 4.89 | 3.98 | 3.53 | 3.47 |
| SWB subset | Hub98 | m & v | 47.9 | 45.2 | 45.1 | 45.0 |
| | Hub01 | m & v | 37.2 | 34.1 | 34.0 | 33.8 |
| SWB *h5train00* | Hub98 | m & v | 41.3 | 39.5 | 39.4 | 39.2 |
| | Hub01 | m & v | 31.5 | 29.5 | 29.6 | 29.3 |

$\rho = 20$, while a larger value of $\alpha$ in this case may make the MMI objective function grow more slowly. On the other hand, when the size of trust region, i.e., $\rho$, becomes larger, we need to penalize the auxiliary function more to make it a lower bound of the original objective function in a wider range. Therefore, a larger value is required for $\alpha$ to achieve a better learning curve of the objective function. Finally, we also summarize the best recognition performance in word error rate (WER) of different optimization methods in Table II. The results show that in the WSJ-5k task both TR-based methods can achieve much lower WER than the EBW method, and the proposed bounded TR method yields slightly better performance than the original TR method without using the penalized term.

Next, we choose a fixed set of configuration of $\rho$ and $\alpha$ (e.g., $\rho = 20$, $\alpha = 0.1$), to compare performance of updating Gaussian means only versus that of updating both Gaussian means and variances simultaneously. The learning curves of the MMI objective function are shown in Fig. 5. We can see that recognition of updating both Gaussian means and variances significantly outperform that of updating means only for both the EBW method and TR-based methods. When updating both means and variances, the bounded TR method still yields the best learning curve of the objective function. According to the recognition performance summarized in Table II, the TR-based methods, when updating both Gaussian means and variances, can achieve about 0.25% absolute reduction in WER performance comparing with that of updating Gaussian means only. When updating both Gaussian mean and variances, the bounded TR method gives about 30% relative WER reduction over the MLE baseline, roughly 13% relative WER reduction over EBW and about 2% reduction over the original TR method without using penalized term. Moreover, we have also conducted significance tests [15] for TR versus EBW, and bounded TR versus EBW. Results show that the original TR is significantly better than EBW ($\mathrm{p} = 0.035$ for updating means only and $\mathrm{p} = 0.038$ for updating both means and variances), and the bounded TR is also significantly better than EBW ($\mathrm{p} = 0.009$ for updating means only and $\mathrm{p} = 0.012$ for updating both means and variances).

### B. Switchboard 60-Hour Subset Task

In the Switchboard subset task, we randomly choose about 60 hours training data from the full *h5train00* (about 265 hours in total) to build a set of cross-word state-tying triphone HMMs. The used feature vectors are 39-dimension PLP including delta and delta-delta features. Vocal tract length normalization (VTLN) is applied to normalize features across different speakers. We evaluate recognition performance on two separate sets, namely Hub5 *eval98* and *eval01* sets. In our experiments, we use the *eval98* set as the development set to tune various parameters in the algorithms and use the *eval01* set as an independent test set to evaluate the best models. The NIST scoring software [15] has been used to measure word error rates (WER). The WER of the MLE baseline is 47.9% on *eval98* and 37.2% on *eval01*.

In the MMI training, both silence and garbage models are kept unchanged in DT since statistics collected from lattices for them are abnormally large, which severely affects convergence of model training. In these experiments, we choose to use

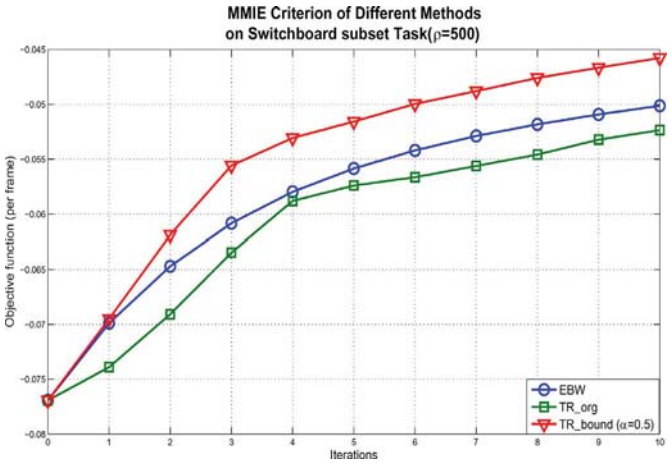


Fig. 6. Comparison of MMI objective function for different optimization methods, trained on Switchboard 60-hour subset and evaluated on eval98. ($\rho = 500$).

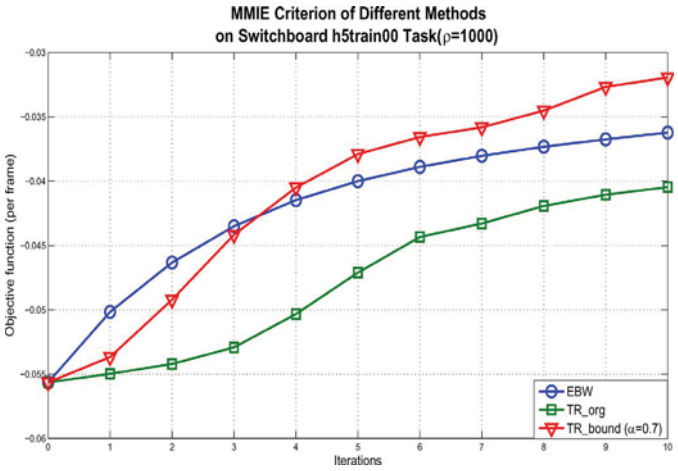


Fig. 7. Comparison of MMI objective function for different optimization methods, trained on Switchboard *h5train00* set and evaluated on eval98.

the best parameter values $\rho = 500$, $\alpha = 0.5$ for the TR-based methods and update both Gaussian means and variances at the same time. From Fig. 6, we can see that the original TR method is not doing as well as the EBW method in term of maximizing the MMI objective function in this task while the bounded TR method still yields the best learning curve among all and it converges to much better value than other methods. From the recognition results in Table II, we can also see that both TR-based methods achieve slightly better recognition performance than the EBW method on *eval98*, and the bounded TR method gives the best result, which is 45.0% in WER, about 6.1% relative WER reduction over the MLE baseline, and about 0.2% absolute improvement from the conventional EBW method. The same results are also observed on the independent *eval01* test set. Significance tests [15] conducted in this task have shown that the bounded TR method significantly outperforms the EBW method ($\mathrm{p} = 0.012$) in *eval01* but the original TR is not significantly better than EBW in *eval01*.

### C. Switchboard h5train00 Full Set

In this section, we use the full *h5train00* set (about 265 hours in total) as training data, which contains the Switchboard (SWB1) corpus and Call Home English (CHE) data, to build cross-word state-tying triphone HMMs. The used features are still 39-dimension PLPs including delta and delta-delta. In these experiments, VTLN is first applied to normalize features across different speakers and then the SAT technique [1] is applied in the feature space to normalize for each conversion side. We still use the eval98 set as the development set and the standard NIST scoring software is still used to measure word error rates (WER). The WER of the MLE baseline is 41.3%, which is comparable with the best single pass performance reported in this task.

We also keep both silence and garbage models unchanged in the MMI training to avoid the model convergence problem caused by abnormally large statistics collected for these two models from lattices. We choose the best parameters $\rho = 1000$, $\alpha = 0.7$ for the both TR methods, and update both means and variances at the same time. From Fig. 7, we have a similar observation as the SWB subset that the MMI objective function of the original TR method is not as good as that of the EBW method. However, the bounded TR method still gives a much better growing curve that leads to a significantly improved convergence point. In Table II, we also see that both TR methods give slightly better recognition performance than the EBW method on *eval98*. More specifically, the bounded TR method yields the best recognition performance (39.2% in WER), which is about 2.1% absolute improvement from the MLE baseline and about 0.3% absolute gain over the EBW method. Finally, we have evaluated the best models on the independent evaluation set *eval01*. From the results shown in Table II, we can see that the bounded TR method still yields the best performance (29.3% in WER) but the TR method is slightly worse than EBW in this case. Significance tests [15] conducted in this task have also shown that the bounded TR method significantly outperforms the EBW method in a weak sense ($\mathrm{p} = 0.047$) in *eval01* but the original TR is not significantly different from EBW in *eval01*.

## VI. Conclusion

This paper presents two trust region (TR) based parameter optimization methods for MMI-based discriminative training of HMMs in speech recognition. In these methods, we derive an auxiliary function to approximate the original discriminative objective function, and imposes a locality constraint to ensure the auxiliary function serves as a good local approximation as well as a strict lower bound of the objective function during optimization. More importantly, a fast global optimization algorithm proposed in optimization theory can be used to optimize the derived auxiliary functions very efficiently. Experimental results on both WSJ-5k and Switchboard tasks have shown that the proposed trust region (TR) method yields better recognition performance than the conventional EBW method, in terms of both discriminative criterion as well as recognition performance. In this work, we only consider discriminative training based on maximum mutual information (MMI) training. As one direction of possible future work, it is straightforward to extend the propose TR methods to deal with other discriminative training criteria in speech recognition, such as MCE, MPE/MWE, and LME.

## APPENDIX
## DERIVATION OF THE GAP FUNCTION BETWEEN Q AND F

In this appendix, we consider to derive the gap function between the original auxiliary function $\mathcal{Q}(\cdot)$ and the MMI objective function $\mathbf{F}_{\mathbf{MMI}}(\cdot)$.

First, we express the MMI objective function as

$$\mathbf{F}_{\mathrm{MMI}}(\Lambda)=\sum_r\Big[\log p(O_r,S_r\mid\Lambda)-\log\sum_{s_r\in\mathcal{M}_r}p(O_r,s_r\mid\Lambda)\Big]. \tag{41}$$

Based on the well-known Bayes' theorem, $p(\mathbf{l}_r \mid O_r,S_r,\Lambda) = p(\mathbf{l}_r,O_r,S_r\mid\Lambda)/p(O_r,S_r\mid\Lambda)$, where $\mathbf{l}_r$ denotes hidden state sequences of $O_r$ in HMM $\Lambda$. The positive term in (41), $\log p(O_r,S_r\mid\Lambda)$, can be represented as

$$\log p(O_r,S_r\mid\Lambda)=\log p(\mathbf{l}_r,O_r,S_r\mid\Lambda)-\log p(\mathbf{l}_r\mid O_r,S_r,\Lambda). \tag{42}$$

If we take expectation on both sides of (42) based on probability distributions of $p(\mathbf{l}_r\mid O_r,S_r,\Lambda^{(n)})$, we have

$$\begin{aligned}&\sum_{\mathbf{l}_r}\log p(O_r,S_r\mid\Lambda)\cdot p(\mathbf{l}_r\mid O_r,S_r,\Lambda^{(n)})\\&=\sum_{\mathbf{l}_r}\log p(\mathbf{l}_r,O_r,S_r\mid\Lambda)\cdot p(\mathbf{l}_r\mid O_r,S_r,\Lambda^{(n)})\\&\quad-\sum_{\mathbf{l}_r}\log p(\mathbf{l}_r\mid O_r,S_r,\Lambda)\cdot p(\mathbf{l}_r\mid O_r,S_r,\Lambda^{(n)}).\end{aligned} \tag{43}$$

It is easy to see that left-hand side of (43) equals to $\log p(O_r,S_r\mid\Lambda)$ because of the sum-to-one property of probability distributions, $p(\mathbf{l}_r\mid O_r,S_r,\Lambda^{(n)})$. Therefore, we can decompose the positive term as follows:

$$\begin{aligned}\log p(O_r,S_r\mid\Lambda)&=\sum_{\mathbf{l}_r}p(\mathbf{l}_r\mid O_r,S_r,\Lambda^{(n)})\\&\quad\cdot\Big[\log p(\mathbf{l}_r,O_r,S_r\mid\Lambda)-\log p(\mathbf{l}_r\mid O_r,S_r,\Lambda)\Big].\end{aligned} \tag{44}$$

Similarly, the same idea can be applied to the negative term in (41), which can be decomposed as follows:

$$\begin{aligned}\log\sum_{s_r\in\mathcal{M}_r}p(O_r,s_r\mid\Lambda)&=\sum_{s_r\in\mathcal{M}_r}\sum_{\mathbf{l}_r}p(\mathbf{l}_r\mid O_r,\mathcal{M}_r,\Lambda^{(n)})\\&\quad\cdot\Big[\log p(\mathbf{l}_r,O_r,s_r\mid\Lambda)-\log p(\mathbf{l}_r\mid O_r,\mathcal{M}_r,\Lambda)\Big].\end{aligned} \tag{45}$$

After substituting (44) and (45) into the MMI objective function in (41), $\mathbf{F}_{\mathrm{MMI}}$ can be presented as follows:

$$\begin{aligned}\mathbf{F}_{\mathrm{MMI}}(\Lambda|\Lambda^{(n)})&=\sum_r\sum_{\mathbf{l}_r}p(\mathbf{l}_r|O_r,S_r,\Lambda^{(n)})\\&\quad\cdot\Big[\log p(\mathbf{l}_r,O_r,S_r|\Lambda)-\log p(\mathbf{l}_r|O_r,S_r,\Lambda)\Big]\\&\quad-\sum_r\sum_{s_r\in\mathcal{M}_r}\sum_{\mathbf{l}_r}p(\mathbf{l}_r|O_r,\mathcal{M}_r,\Lambda^{(n)})\\&\quad\cdot\Big[\log p(\mathbf{l}_r,O_r,s_r|\Lambda)-\log p(\mathbf{l}_r|O_r,\mathcal{M}_r,\Lambda)\Big].\end{aligned} \tag{46}$$

In the MMI training, as shown in Section III-C , we normally construct the following auxiliary function according to the well-know Jensen's inequality as follows:

$$\begin{aligned}\mathcal{Q}(\Lambda|\Lambda^{(n)})&=\sum_r\sum_{\mathbf{l}_r}p(\mathbf{l}_r|O_r,S_r,\Lambda^{(n)})\\&\quad\cdot\Big[\log p(\mathbf{l}_r,O_r,S_r|\Lambda)-\log p(\mathbf{l}_r|O_r,S_r,\Lambda^{(n)})\Big]\\&\quad-\sum_r\sum_{s_r\in\mathcal{M}_r}\sum_{\mathbf{l}_r}p(\mathbf{l}_r|O_r,\mathcal{M}_r,\Lambda^{(n)})\\&\quad\cdot\Big[\log p(\mathbf{l}_r,O_r,s_r|\Lambda)-\log p(\mathbf{l}_r|O_r,\mathcal{M}_r,\Lambda^{(n)})\Big].\end{aligned} \tag{47}$$

It is easy to see that the constant parts w.r.t $\Lambda$, i.e.,

$$\begin{aligned}&-\sum_r\sum_{\mathbf{l}_r}p(\mathbf{l}_r|O_r,S_r,\Lambda^{(n)})\cdot\log p(\mathbf{l}_r|O_r,S_r,\Lambda^{(n)})\\&\quad+\sum_r\sum_{\mathbf{l}_r\in\mathcal{M}_r}p(\mathbf{l}_r|O_r,\mathcal{M}_r,\Lambda^{(n)})\cdot\log p(\mathbf{l}_r|O_r,\mathcal{M}_r,\Lambda^{(n)})\end{aligned}$$

is absorbed into $C_1$ as in (16).

Finally, based on (46) and (47), the gap function $\mathcal{G}(\Lambda|\Lambda^{(n)})$ can be calculated as follows:

$$\begin{aligned}&\mathcal{G}(\Lambda|\Lambda^{(n)})\\&=\mathcal{Q}(\Lambda|\Lambda^{(n)})-\mathbf{F}_{\mathrm{MMI}}(\Lambda|\Lambda^{(n)})\\&=\sum_r\sum_{\mathbf{l}_r}p(\mathbf{l}_r|O_r,S_r,\Lambda^{(n)})\log\frac{p(\mathbf{l}_r\mid O_r,S_r,\Lambda)}{p(\mathbf{l}_r\mid O_r,S_r,\Lambda^{(n)})}\\&\quad-\sum_r\sum_{\mathbf{l}_r\in\mathcal{M}_r}p(\mathbf{l}_r|O_r,\mathcal{M}_r,\Lambda^{(n)})\log\frac{p(\mathbf{l}_r\mid O_r,\mathcal{M}_r,\Lambda)}{p(\mathbf{l}_r\mid O_r,\mathcal{M}_r,\Lambda^{(n)})}.\end{aligned} \tag{48}$$

## REFERENCES


[1] T. Anastasakos, J. McDonough, R. Schwarts, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. Int. Conf. Spoken Lang. Process.*, 1996, pp. 1137–1140.

[2] 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," in *Proc. ICASSP'86*, Tokyo, Japan, 1986, pp. 49–52.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *J. R. Statist. Soc. B*, vol. 39, pp. 1–38, 1977.

[4] X. He, L. Deng, and W. Chou, "Discriminative learning in sequential pattern recognition: A unifying review for optimization-based speech recognition," *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 14–14–36, Sep. 2008.

[5] H. Jiang, X. Li, and C.-J. Liu, "Large margin hidden Markov models for speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1584–1595, Sep. 2006.

[6] H. Jiang and X. Li, *A General Approximation-Optimization Approach to Large Margin Estimation of HMMs*. Vienna, Austria: I-TECH Education and Publishing, 2007, Robust Speech Recognition and Understanding, ch. 7, pp. 103–120.

[7] H. Jiang and X. Li, "Parameter estimation of statistical models using convex optimization: An advanced method of discriminative training for speech and language processing," *IEEE Signal Process. Mag.* , vol. 27, no. 3, pp. 115–127, May 2010.

[8] H. Jiang, "Discriminative training of HMMs for automatic speech recognition: A survey," *Comput. Speech Lang.*, vol. 24, no. 4, pp. 589–608, Oct. 2010.

[9] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 257–265, May 1997.

[10] J. L. M. Yuan and C. -. Lee, "Soft margin estimation of hidden Markov model parameters," in *Proc. Interspeech'06*, 2006, pp. 2422–2425.
[11] 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. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 900–909, Jan. 2008.
[12] P. Liu and F. Soong, "A quadratic optimization approach to discriminative training of CDHMMs," *IEEE Signal Process. Lett.*, vol. 16, no. 1, pp. 149–152, Mar. 2009.
[13] E. McDermott, T. Hazen, J. L. Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large-vocabulary speech recognition using minimum classification error," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 203–223, Jan. 2007.
[14] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York: Springer, 2006.
[15] D. S. Pallet, W. M. Fisher, and J. G. Fiscus, "Tools for the analysis of benchmark speech recognition tests," in *Proc. ICASSP'90*, 1990.
[16] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP'02*, Orlando, FL, 2002, pp. 105–108.
[17] D. Povey, "Discriminative Training for Large Vocabulary Speech Recognition," Ph.D. dissertation, Cambridge Univ., Cambridge, U.K., 2004.
[18] F. Sha and L. K. Saul, "Large margin hidden Markov models for automatic speech recognition," in *Proc. NIPS'07*, 2007.
[19] D. C. Sorensen, "Newton's method with a model trust region modification," *SIAM J. Numer. Anal.*, vol. 19, no. 2, pp. 409–426, Apr. 1982.
[20] Z.-J. Yan, C. Liu, Y. Hu, and H. Jiang, "A trust region based optimization for maximum mutual information estimation of HMMs in speech recognition," in *Proc. ICASSP09*, 2009, pp. 4521–4524.
[21] D. Yu, L. Deng, X. He, and A. Acero, "Large-margin minimum classification error training for large-scale speech recognition tasks," in *Proc. ICASSP07*, 2007, pp. 1137–1140.
[22] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Comput. Speech Lang.*, vol. 16, no. 1, pp. 25–47, Jan. 2002.

**Cong Liu** received the Ph.D. degree from the University of Science and Technology of China (USTC), Hefei, in July 2010.

From September 2004 to July 2010, he was with the iFLYTEK Speech Lab, USTC. From July 2006 to December 2006, he was a visiting student at Microsoft Research Asia, Beijing, China. Since July 2010, he has been a Researcher with iFLYTEK Research, Hefei. His research interests include speech recognition and speaker recognition, especially in acoustic modeling and discriminative training.

**Yu Hu** received the B.Eng., M.Eng., and Ph.D. degrees from the University of Science and Technology of China (USTC), Hefei, China, in 2000, 2003, and 2009, respectively, all in electrical engineering.

In 1999, he became a Research Engineer with iFlytek Ltd. as a cofounder, working on Mandarin speech synthesis and speech prosody analysis. He was one of researchers who built the first few generations of iFlytek Mandarin speech synthesis engines. Since 2004, his research interest has changed to robust speech recognition, and began to work on the iFlytek Mandarin speech recognition system. He is currently the Director of iFlytek Research and working on speech recognition over mobile internet.

**Li-Rong Dai** was born in China in 1962. He received the B.S. degree in electrical engineering from Xidian University, Xi'an, China, and the M.S. degree from Hefei University of Technology, Hefei, China, in 1983 and 1986, respectively, and the Ph.D. degree in signal and information processing from the University of Science and Technology of China (USTC), Hefei, in 1997.

He joined USTC in 1993. He is currently a Professor in the School of Information Science and Technology, USTC. His current research interests include speech synthesis, speaker and language recognition, speech recognition, digital signal processing, voice search technology, machine learning, and pattern recognition. He has published more than 50 papers in these areas.

**Hui Jiang** (M'00–SM'11) received the B.Eng. and M.Eng. degrees from the University of Science and Technology of China (USTC), Hefei, in 1992 and 1994, respectively, and the Ph.D. degree from the University of Tokyo, Tokyo, Japan, in September 1998, all in electrical engineering.

From October 1998 to April 1999, he was a Researcher in the University of Tokyo. From April 1999 to June 2000, he was with Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, as a Postdoctoral Fellow. From 2000 to 2002, he worked in Dialogue Systems Research, Multimedia Communication Research Lab, Bell Labs, Lucent Technologies Inc., Murray Hill, NJ. He joined the Department of Computer Science and Engineering, York University, Toronto, ON, Canada, as an Assistant Professor in fall 2002 and was promoted to Associate Professor in 2007. His current research interests include speech and audio processing, machine learning, statistical data modeling, and bioinformatics, especially discriminative training, robustness, noise reduction, utterance verification, and confidence measures.

Dr. Jiang has served as an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING since 2009.