A Constrained Line Search Optimization Method for Discriminative Training of HMMs

Peng Liu, Cong Liu, Hui Jiang, *Member, IEEE*, Frank Soong, *Senior Member, IEEE*, and Ren-Hua Wang, *Member, IEEE*

***Abstract*—In this paper, we propose a novel optimization algorithm called *constrained line search (CLS)* for discriminative training (DT) of Gaussian mixture continuous density hidden Markov model (CDHMM) in speech recognition. The CLS method is formulated under a general framework for optimizing any discriminative objective functions including maximum mutual information (MMI), minimum classification error (MCE), minimum phone error (MPE)/minimum word error (MWE), etc. In this method, discriminative training of HMM is first cast as a constrained optimization problem, where Kullback–Leibler divergence (KLD) between models is explicitly imposed as a constraint during optimization. Based upon the idea of line search, we show that a simple formula of HMM parameters can be found by constraining the KLD between HMM of two successive iterations in an quadratic form. The proposed CLS method can be applied to optimize all model parameters in Gaussian mixture CDHMMs, including means, covariances, and mixture weights. We have investigated the proposed CLS approach on several benchmark speech recognition databases, including TIDIGITS, Resource Management (RM), and Switchboard. Experimental results show that the new CLS optimization method consistently outperforms the conventional EBW method in both recognition performance and convergence behavior.**

***Index Terms*—Discriminative training (DT), optimization algorithm, line search, Kullback–Leibler divergence (KLD).**

I. Introduction

In THE past few decades, discriminative training (DT) has been a very active research area in automatic speech recognition (ASR). Most discriminative training methods have been formulated to estimate parameters of Gaussian mixture continuous density hidden Markov models (CDHMM) in different speech recognition tasks, ranging from small vocabulary, isolated word recognition to large vocabulary, continuous speech recognition tasks. Discriminative training of CDHMMs is a typical optimization problem, where an objective function is usually optimized in an iterative manner. Popular DT criteria includes maximum mutual information (MMI)[1], minimum classification error (MCE) [2]–[4], minimum word or phone error (MWE or MPE) [5], minimum divergence (MD) [6], etc. Once the objective function is chosen, an effective algorithm is used to optimize the objective function by adjusting CDHMM parameters. In speech recognition, various algorithms have been proposed to optimize the objective function, including the generalized probabilistic descent (GPD) algorithm based on the first-order gradient descent, the approximate second-order Quickprop algorithm, and the extended Baum–Welch (EBW) algorithm, etc. The GPD and Quickprop methods are mainly used for optimizing the MCE objective function. The EBW method has been initially proposed for optimizing a rational objective function and later extended to Gaussian mixture CDHMMs for the MMI [7] and MPE (or MWE) [8] objective functions. Recently, the EBW method has also been generalized to optimize the MCE objective function [9] and the MD objective function [6]. Nowadays, the EBW method has been widely accepted for discriminative training because it is relatively easy to implement the EBW algorithm on word graphs for large scale ASR tasks and it has been demonstrated that the EBW algorithm performs quite well on many ASR tasks. Generally speaking, these optimization methods attempt to search for a nearby locally optimal point of objective functions from any initial point according to both a search direction and a step size. Normally, the search direction is locally computed based on the first-order derivative (such as gradient), but the step size must be empirically determined in practice. As a result, the performance of these optimization methods highly depends on the location of the initial point and the property of objective functions. If the derived objective function is highly nonlinear, jagged, and nonconvex in nature, it is usually difficult to optimize it effectively with any simple optimization algorithm. In speech recognition, the major difficulties of discriminative training lie in high dimensionality of model parameters and highly complex nature of the derived objective functions. In practice, some heuristic methods have been used to smooth the objective functions to make them optimizable, such as the so-called acoustic scaling in [1] and the use of a smooth sigmoid function in MCE [2], [10] and so on.

In this paper, we propose a novel optimization method, called *constrained line search (CLS)*, for discriminative training of Gaussian mixture CDHMMs. As a general optimization method put forward under a unified framework, the proposed constrained line search method is capable of optimizing various popular DT objective functions in speech recognition, including: MMI, MCE, MPE (or MWE), MD, etc. And a simple closed-form solution can be derived to efficiently update means, covariances, and mixture weights of Gaussian mixture CDHMMs. In this paper, we first cast the discriminative training of CDHMMs as a constrained optimization problem, where a constraint is explicitly imposed for DT based on the

Manuscript received October 14, 2007; revised March 31, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerhard Rigoll.

P. Liu and F. Soong are with Microsoft Research Asia, Beijing 100080, China (e-mail: pengliu@microsoft.com; frankkps@microsoft.com).

C. Liu and R.-H. Wang are with the University of Science and Technology of China, Hefei 230026 , China (e-mail: yylhbt@mail.ustc.edu.cn; rhw@ustc.edu.cn).

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

Digital Object Identifier 10.1109/TASL.2008.925882

Kullback–Leibler divergence (KLD) [11] between CDHMMs between two successive iteration. The constraint is motivated by the fact that all collected estimation statistics are only reliable in a close neighborhood of the original model. Under this constraint, the objective function can be approximated as a smooth quadratic function of CDHMM parameters, and its sole critical point, if existing, can be easily obtained by vanishing its derivative to zero. Then, a novel algorithm called *constrained line search* (CLS) is proposed to solve the constrained optimization problem. Subject to the KLD constraint, the line search is performed either along a line segment joining the initial model parameters and the critical point of the smoothed objective function, if the critical point exists, or along gradient direction of the objective function at the initial point, if the critical point does not exist. As we will show, a closed-form solution can be derived as long as we can formulate or approximate the KLD constraint as quadratic constraint.

In this paper, the proposed CLS method has been used to optimize the MMI objective function as well as other DT objective functions in several speech recognition tasks, including connected digit string recognition using the TIDIGITS database [13], the resource management (RM) task [14], and large-vocabulary continuous speech recognition on the Switchboard task [15]. The experimental results clearly show that the proposed CLS algorithm outperforms the popular EBW method in all evaluated ASR tasks in terms of final recognition performance and convergence behavior.

The remainder of the paper is organized as follows: In Section II, we first formulate discriminative training as a KLD constrained optimization problem under a general framework. In Section III, we discuss how to simplify and approximate the KLD constraints into a quadratic form. In Section IV, we describe the constrained line search (CLS) optimization method in details and derive closed-form solutions to update all CDHMM parameters for CLS. In Section V, we examine the proposed CLS method on several standard speech recognition tasks and report and discuss experimental results. Finally, we conclude the paper with our findings and discuss about future works in the Conclusion

II. Formulation of Discriminative Training as Constrained Optimization

A. Criteria of Discriminative Training

We assume that an acoustic model set $\mathbf{\Lambda}$ consists of many individual Gaussian mixture CDHMMs, each is represented as a triple $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$, where $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_N\}$ is the initial state distribution and N is the number of states in HMM, $\boldsymbol{A} = \{a_{ij}\}_{N\times N}$ is transition matrix, and $\boldsymbol{B}$ is state output distribution set, consisting of Gaussian mixture distributions for all states: $b_i(\boldsymbol{x}) = \sum_{k=1}^{K} \omega_{ik} \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{ik}, \mathbf{\Sigma}_{ik})$, where $\boldsymbol{\theta}_i = \{\boldsymbol{\omega}_i, \boldsymbol{\mu}_{ik}, \mathbf{\Sigma}_{ik} \,|\, 1 \le k \le K\}$ with K standing for the number of Gaussian mixture components in state $i (1 \le i \le N)$, and $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{\Sigma})$ for a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$.

For a training utterance $\boldsymbol{X}$ and its corresponding transcription W, we first consider how to compute acoustic model score $p(\boldsymbol{X} \,|\, W)$ based on the composite HMM $\boldsymbol{\lambda}_W$ of W. Suppose $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$, let $\boldsymbol{s} = \{s_1, s_2, \ldots, s_T\}$ be any possible state sequence, and $\boldsymbol{l} = \{l_1, l_2, \ldots, l_T\}$ be the associated sequence of the mixture component labels. The likelihood $p(\boldsymbol{X} \,|\, W)$ is computed as

$$p(\boldsymbol{X} \,|\, W) = \sum_{\boldsymbol{s}} \sum_{\boldsymbol{l}} \left\{ \pi_{s_1} \prod_{t=2}^{T} a_{s_{t-1}s_t} \prod_{t=1}^{T} \omega_{s_t l_t} \mathcal{N} \times \left(\boldsymbol{x}_t; \boldsymbol{\mu}_{s_t l_t}, \mathbf{\Sigma}_{s_t l_t}\right) \right\} \quad (1)$$

where summations are taken over all possible state sequences $\boldsymbol{s}$ and mixture labels $\boldsymbol{l}$.

Assume that the whole training set consists of R different training utterances $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_R\}$ along with their corresponding transcriptions, denoted as $\{W_1, W_2, \ldots, W_R\}$. As shown in [16], objective functions of CDHMMs for various discriminative training criteria can be formulated in the following general form, as shown by (2) at the bottom of the page, where $0 < \kappa \le 1$ is the acoustic scaling factor. (It is remarkable that for MCE there is not an explicit acoustic scaling factor. However, the smoothing factor of α [2] in MCE is essentially playing a similar role, and the MCE approach can also be represented in a posterior form [16].) And $\mathcal{M}_r$ stands for all competing string hypotheses of utterance $\boldsymbol{X}_r$ which is compactly approximated by a word lattice generated in Viterbi decoding, $f(\cdot)$ is a mapping function to transform the objective function, and $G(W, W_r)$ is the *gain function* to measure dissimilarity between reference W_r and a hypothesis W. The mapping function $f(\cdot)$ and the gain function $G(W, W_r)$ can take different functional forms in various discriminative training criteria, as listed in Table I. In this study, we assume that language model score $p(W)$ is fixed.

B. Constrained Optimization for Discriminative Training

After substituting (1) into (2), the general DT objective function $\mathcal{F}(\mathbf{\Lambda})$ becomes a complicated, highly nonlinear function, which is difficult to optimize directly. Therefore, we normally make the following assumptions: 1) competing recognition hypothesis space $\mathcal{M}_r$ remains unchanged throughout the whole optimization; and 2) the estimation statistics, including state occupancies and Gaussian kernel occupancies, remains unchanged

$$\begin{aligned} \mathcal{F}(\mathbf{\Lambda}) &= p(\mathbf{\Lambda} | \{\boldsymbol{X}_r, W_r, \mathcal{M}_r\}_{r=1}^{R}, f, \kappa, G) \\ &= \frac{1}{R} \sum_{r=1}^{R} f\left(\log \left[\frac{\sum_{W \in \mathcal{M}_r} p^{\kappa}(\boldsymbol{X}_r \,|\, W_r) p(W_r) G(W, W_r)}{\sum_{W' \in \mathcal{M}_r} p^{\kappa}(\boldsymbol{X}_r \,|\, W') p(W')} \right]^{\frac{1}{\kappa}} \right) \end{aligned} \quad (2)$$

TABLE I
OBJECTIVE FUNCTIONS FOR VARIOUS DISCRIMINATIVE TRAINING CRITERIA ($\delta(\cdot)$: *KRONECKER DELTA FUNCTION*; $|\cdot|$: *NUMBER OF SYMBOLS IN A STRING*; $\mathrm{LEV}(\cdot \parallel \cdot)$: *LEVENSHITINE DISTANCE*; $\mathcal{D}(\cdot \parallel \cdot)$: *KLD*)

Criterion	Mapping function $f(z)$	Alternative word sequences $\mathcal{M}_r$	Gain function $G(W, W_r)$
MMI	z	All(recognized)	$\delta(W, W_r)$
MWE	e^z	All(recognized)	$\lvert W_r\rvert - \mathrm{LEV}(\mathrm{W} \parallel \mathrm{W_r})$
MD	e^z	All(recognized)	$-\mathcal{D}(W \parallel W_r)$

during optimization. We also use a sufficiently small scaling factor $\kappa(\kappa \ll 1)$ to smooth the original objective function. Because of these, it makes sense to explicitly impose a constraint that HMM model parameters $\mathbf{\Lambda}$ do not deviate too much from their initial values, $\mathbf{\Lambda}^0$. This constraint ensures that all of the above assumptions remain valid during optimization since the initial models $\mathbf{\Lambda}^0$ have been used to generate all word lattices $\{\mathcal{M}_r\}$ and corresponding statistics from the training data for optimization.

The constraint can be quantitatively defined based on the KLD between models. Therefore, given an initial model set $\mathbf{\Lambda}^0$, we formulate discriminative training of CDHMMs as the following constrained maximization problem:

$$\mathbf{\Lambda}^* = \arg\max_{\mathbf{\Lambda}} \mathcal{F}(\mathbf{\Lambda}) \tag{3}$$

$$\text{subject to } \mathcal{D}(\mathbf{\Lambda} \| \mathbf{\Lambda}^0) \le \rho^2 \tag{4}$$

where $\mathcal{D}(\mathbf{\Lambda} \,\|\, \mathbf{\Lambda^0})$ is the KLD between $\mathbf{\Lambda}$ and $\mathbf{\Lambda}^0$, and $\rho > 0$ is a preset constant to control the search range. The constraint in (4) specifies a *trust region* of optimization. Note that theoretical analysis of discriminative training in [17] and [18] also supports to use such a constraint in discriminative training.

III. KLD CONSTRAINTS FOR CDHMMS

First of all, we formulate the KLD-based model constraint in (4) for different CDHMM parameters. As long as we approximate the KLD constraints in a quadratic form, we will be able to derive a simple closed-form solution for updating CDHMM parameters.

A. Constraint Decomposition for Gaussian Mixtures

Assume the whole model set $\mathbf{\Lambda}$ consists of many CDHMMs $\boldsymbol{\lambda}(\boldsymbol{\lambda} \in \mathbf{\Lambda})$, the overall KLD constraint in (4) can be ensured by many local constraints for all individual CDHMMs as $\mathcal{D}(\boldsymbol{\lambda} \,\|\, \boldsymbol{\lambda}^0) \le \rho^2 (\boldsymbol{\lambda} \in \mathbf{\Lambda})$.

Furthermore, $\mathcal{D}(\boldsymbol{\lambda} \,\|\, \boldsymbol{\lambda}^0)$ can be decomposed into its Gaussian components

$$\begin{aligned}\mathcal{D}(\boldsymbol{\lambda} \,\|\, \boldsymbol{\lambda}^0) &\le \sum_{i=1}^{N} \mathcal{D}\left(\boldsymbol{\theta}_i \,\|\, \boldsymbol{\theta}_i^0\right) \\ &\le \sum_{i=1}^{N} \left[\mathcal{D}\left(\boldsymbol{\omega}_i \,\|\, \boldsymbol{\omega}_i^0\right) + \sum_{k=1}^{K} \omega_{ik} \right. \\ &\qquad \left. \times \mathcal{D}\left(\mathcal{N}\left(\boldsymbol{\mu}_{ik}, \mathbf{\Sigma}_{ik}\right) \| \mathcal{N}\left(\boldsymbol{\mu}_{ik}^0, \mathbf{\Sigma}_{ik}^0\right)\right) \right]\end{aligned} \tag{5}$$

where $\boldsymbol{\omega}_i = (\omega_{i1}, \omega_{i2}, \ldots, \omega_{iK})'$ denotes all Gaussian mixture weights, $\mathcal{D}(\boldsymbol{\omega}_i \,\|\, \boldsymbol{\omega}_i^0)$ denotes KLD between $\boldsymbol{\omega}_i$ and its initial value, and $\mathcal{D}(\mathcal{N}(\boldsymbol{\mu}_{ik}, \mathbf{\Sigma}_{ik}) \,\|\, \mathcal{N}(\boldsymbol{\mu}_{ik}^0, \mathbf{\Sigma}_{ik}^0))$ denotes the KLD between two multivariate Gaussian distributions. By definition, they can be computed as

$$\mathcal{D}\left(\boldsymbol{\omega}_i \,\|\, \boldsymbol{\omega}_i^0\right) = \sum_{k=1}^{K} \omega_{ik} \left(\log \frac{\omega_{ik}}{\omega_{ik}^0}\right) = \boldsymbol{\omega}_i' \left(\log \boldsymbol{\omega}_i - \log \boldsymbol{\omega}_i^0\right) \tag{6}$$

$$\begin{aligned}&\mathcal{D}\left(\mathcal{N}(\boldsymbol{\mu}_{ik}, \mathbf{\Sigma}_{ik}) \,\|\, \mathcal{N}\left(\boldsymbol{\mu}_{ik}^0, \mathbf{\Sigma}_{ik}^0\right)\right) \\ &= \frac{1}{2}\left[\mathrm{tr}\left[\left(\mathbf{\Sigma}_{ik}^0\right)^{-1} \mathbf{\Sigma}_{ik}\right] + \log \frac{\left|\mathbf{\Sigma}_{ik}^0\right|}{\left|\mathbf{\Sigma}_{ik}\right|} \right. \\ &\qquad \left. + \left(\boldsymbol{\mu}_{ik} - \boldsymbol{\mu}_{ik}^0\right)' \left(\mathbf{\Sigma}_{ik}^0\right)^{-1} \left(\boldsymbol{\mu}_{ik} - \boldsymbol{\mu}_{ik}^0\right) - D \right]\end{aligned} \tag{7}$$

where D is the dimension of each Gaussian kernel.

We can further break down the constraint separately in (7) for Gaussian means, covariance, and weights. The overall constraint is decomposed into the following independent parts for different model parameters:

$$\begin{aligned}\mathcal{D}\left(\boldsymbol{\mu}_{ik} \,\|\, \boldsymbol{\mu}_{ik}^0\right) &= \left(\boldsymbol{\mu}_{ik} - \boldsymbol{\mu}_{ik}^0\right)' \left(\mathbf{\Sigma}_{ik}^0\right)^{-1} \\ &\quad \times \left(\boldsymbol{\mu}_{ik} - \boldsymbol{\mu}_{ik}^0\right) \le \rho^2\end{aligned} \tag{8}$$

$$\begin{aligned}\mathcal{D}\left(\mathbf{\Sigma}_{ik} \,\|\, \mathbf{\Sigma}_{ik}^0\right) &= \mathrm{tr}\left[\left(\mathbf{\Sigma}_{ik}^0\right)^{-1} \mathbf{\Sigma}_{ik}\right] \\ &\quad + \log \frac{\left|\mathbf{\Sigma}_{ik}^0\right|}{\left|\mathbf{\Sigma}_{ik}\right|} - D \le \rho^2\end{aligned} \tag{9}$$

$$\mathcal{D}\left(\boldsymbol{\omega}_i \,\|\, \boldsymbol{\omega}_i^0\right) = \boldsymbol{\omega}_i' \cdot \left(\log \boldsymbol{\omega}_i - \log \boldsymbol{\omega}_i^0\right) \le \rho^2 \tag{10}$$

where ρ is a parameter to control the trust region of CDHMM model parameters.

Obviously, the constraints of Gaussian mean vectors follow an quadratic form which can be represented as

$$\begin{aligned}\mathcal{D}\left(\boldsymbol{\mu}_{ik} \,\|\, \boldsymbol{\mu}_{ik}^0\right) &= \left(\boldsymbol{\mu}_{ik} - \boldsymbol{\mu}_{ik}^0\right)' \left(\mathbf{\Sigma}_{ik}^0\right)^{-1} \left(\boldsymbol{\mu}_{ik} - \boldsymbol{\mu}_{ik}^0\right) \\ &\equiv Q\left(\boldsymbol{\mu}_{ik} - \boldsymbol{\mu}_{ik}^0, \mathbf{\Sigma}_{ik}^0\right)\end{aligned} \tag{11}$$

where $Q(\boldsymbol{\mu}, \mathbf{\Sigma})$ stands for a standard quadratic form with a positive-definite matrix $\mathbf{\Sigma}$.

B. Quadratic Approximation for KLD Constraints of Covariance and Mixture Weights

As we will show in Section IV, given an quadratic form constraint, (3) can be easily solved in a closed-form. In this section, we consider to approximate the constraints for covariances and weights into quadratic forms based on the Taylor series.

In this paper, we assume all covariance matrices $\mathbf{\Sigma}_{ik}$ are diagonal: $\mathbf{\Sigma}_{ik} = \mathrm{diag}(\sigma_{ik1}^2, \ldots, \sigma_{ikD}^2)$. For computational convenience, we represent each diagonal covariance matrix as a vector in the logarithm domain: $\boldsymbol{\sigma}_{ik} = (\log \sigma_{ik1}^2, \ldots, \log \sigma_{ikD}^2)'$. Then, we have

$$\begin{aligned}&\mathcal{D}\left(\mathbf{\Sigma}_{ik} \,\|\, \mathbf{\Sigma}_{ik}^0\right) \\ &= \mathrm{tr}\left[\mathbf{\Sigma}_{ik}\left(\mathbf{\Sigma}_{ik}^0\right)^{-1}\right] + \log \frac{\left|\mathbf{\Sigma}_{ik}^0\right|}{\left|\mathbf{\Sigma}_{ik}\right|} - D \\ &= \sum_{d=1}^{D} (e^{y_{ikd}} - y_{ikd} - 1) \approx \sum_{d=1}^{D} y_{ikd}^2/2 \\ &= \left(\boldsymbol{\sigma}_{ik} - \boldsymbol{\sigma}_{ik}^0\right)' \left(\boldsymbol{\sigma}_{ik} - \boldsymbol{\sigma}_{ik}^0\right)/2 \\ &\equiv Q\left(\boldsymbol{\sigma}_{ik} - \boldsymbol{\sigma}_{ik}^0, \boldsymbol{I}\right)/2\end{aligned} \tag{12}$$

where D is the feature dimension of HMMs, and we denote $y_{ikd} = \log(\sigma_{ikd}/\sigma_{ikd}^0)^2$ and we have used the second-order

Taylor series to approximate exponential function as $e^y - y - 1 \approx y^2/2$.

As for Gaussian mixture weights $\boldsymbol{\omega}_i$, we denote $z_{ik} = \omega_{ik}/\omega_{ik}^0$. If we adopt the Taylor series approximation $\log z \approx z - 1$, we have

$$\begin{aligned}\mathcal{D}(\boldsymbol{\omega}_i \,\|\, \tilde{\boldsymbol{\omega}}_i) &= \boldsymbol{\omega}_i' \cdot (\log \boldsymbol{\omega}_i - \log \hat{\boldsymbol{\omega}}_i) \\ &= \sum_{k=1}^{K} \omega_{ik} \log z_{ik} \approx \sum_{k=1}^{K} \omega_{ik}(z_{ik} - 1) \\ &= \sum_{k=1}^{K} \frac{\omega_{ik}\omega_{ik} - 2\omega_{ik}^0\omega_{ik} + \omega_{ik}^0\omega_{ik}^0}{\omega_{ik}^0} \\ &= \left(\boldsymbol{\omega}_i - \boldsymbol{\omega}_i^0\right)' \left(\boldsymbol{\Pi}_i^0\right)^{-1} \left(\boldsymbol{\omega}_i - \boldsymbol{\omega}_i^0\right) \\ &\equiv Q\left(\boldsymbol{\omega}_i - \boldsymbol{\omega}_i^0, \boldsymbol{\Pi}_i^0\right) \end{aligned} \tag{13}$$

where $\boldsymbol{\Pi}_i^0 = \operatorname{diag}(\omega_{i1}^0, \ldots, \omega_{iK}^0)$ is a $K \times K$ diagonal positive-definite matrix. In addition to the above quadratic constraint, mixture weights $\boldsymbol{\omega}_i$ must satisfy an affine constraint $\sum_{k=1}^{K} \omega_{ik} = 1$. Note that we have explicitly applied the stochastic constraints $\sum_{k=1}^{K} \omega_{ik} = \sum_{k=1}^{K} \omega_{ik}^0 = 1$ to derive the approximation in (13).

In summary, we approximate the original KLD-based model constraints by the following positive-definite quadratic terms for all model parameters

$$\begin{cases} Q\left(\boldsymbol{\mu}_{ik} - \boldsymbol{\mu}_{ik}^0, \boldsymbol{\Sigma}_{ik}^0\right) \le \rho^2 & (1 \le i \le N) \quad (1 \le k \le K) \\ Q(\boldsymbol{\sigma}_{ik} - \boldsymbol{\sigma}_{ik}^0, \boldsymbol{I}) \le \rho^2 & (1 \le i \le N) \quad (1 \le k \le K) \\ Q\left(\boldsymbol{\omega}_i - \boldsymbol{\omega}_i^0, \boldsymbol{\Pi}_i^0\right) \le \rho^2 & (1 \le i \le N). \end{cases} \tag{14}$$

IV. Constrained Line Search

In this section, we consider how to solve the constrained optimization problem in (3) and (4) and derive closed-form solutions for updating Gaussian mixture based on line search.

If we substitute $\boldsymbol{\mu}_{ik}, \boldsymbol{\sigma}_{ik}$, or $\boldsymbol{\omega}_i$ to (34) in Appendix A in place of $\boldsymbol{\lambda}_{ik}$, we can derive partial derivatives of $\mathcal{F}(\boldsymbol{\Lambda})$ w.r.t. means $\boldsymbol{\mu}_{ik}$, variances $\boldsymbol{\sigma}_{ik}$, and weight $\boldsymbol{\omega}_i$ of each Gaussian mixture component, respectively

$$\nabla \mathcal{F}(\boldsymbol{\mu}_{ik}) = \boldsymbol{\Sigma}_{ik}^{-1}[\mathcal{O}_{ik}(\boldsymbol{x}) - \mathcal{O}_{ik}(1)\boldsymbol{\mu}_{ik}] \tag{15}$$

$$\begin{aligned}\nabla \mathcal{F}(\boldsymbol{\sigma}_{ik}) = \frac{\boldsymbol{\Sigma}_{ik}^{-1}}{2\mathcal{O}_{ik}(1)} &\left[\mathcal{O}_{ik}(1)\mathcal{O}_{ik}(\boldsymbol{x}^2)\right. \\ &\left. - \mathcal{O}_{ik}^2(\boldsymbol{x}) - \mathcal{O}_{ik}^2(1)\exp(\boldsymbol{\sigma}_{ik})\right] \end{aligned} \tag{16}$$

$$\nabla \mathcal{F}(\boldsymbol{\omega}_i) = \boldsymbol{\Pi}_i^{-1} \cdot [\mathcal{O}_{i1}(1), \ldots, \mathcal{O}_{iK}(1)]' \tag{17}$$

where we denote the following statistics for the zeroth-, first-, and second-order moments

$$\mathcal{O}_{ik}(1) = \sum_{r=1}^{R} \sum_{W \in \mathcal{M}_r} C_r(W) \cdot \sum_{t=1}^{T} \gamma_{ik}^W(r,t) \tag{18}$$

$$\mathcal{O}_{ik}(\boldsymbol{x}) = \sum_{r=1}^{R} \sum_{W \in \mathcal{M}_r} C_r(W) \cdot \sum_{t=1}^{T} \gamma_{ik}^W(r,t) \cdot \boldsymbol{x}_{rt} \tag{19}$$

$$\mathcal{O}_{ik}(\boldsymbol{x}^2) = \sum_{r=1}^{R} \sum_{W \in \mathcal{M}_r} C_r(W) \cdot \sum_{t=1}^{T} \gamma_{ik}^W(r,t) \cdot \boldsymbol{x}_{rt}^2 \tag{20}$$

As discussed in Appendix A, if the acoustic scaling factor κ is sufficiently small ($\kappa \ll 1$), we can approximately treat $C_r(W)$ and $\gamma_{ik}^W(r,t)$ (see Appendix A for their definitions) as constants which are independent of CDHMM parameters or slowly changing with respect to model parameters. As a result, all the statistics given above can be treated as constants, and the objective function $\mathcal{F}(\boldsymbol{\Lambda})$ becomes a smooth function so that its unique critical point can be obtained by setting its derivative to zero, i.e., $\nabla \mathcal{F}(\boldsymbol{\Lambda}) = 0$. After solving the equations: $\nabla \mathcal{F}(\boldsymbol{\mu}_{ik}) = 0, \nabla \mathcal{F}(\boldsymbol{\sigma}_{ik}) = 0$, we can easily derive the critical point of the above smoothed objective for Gaussian mean and variances. For Gaussian weights, we use Lagrange multipliers to obtain the critical point subject to the constraint of $\sum_k \omega_{ik} = 1$. The results are shown as follows:

$$\hat{\boldsymbol{\mu}}_{ik} = \frac{\mathcal{O}_{ik}(\boldsymbol{x})}{\mathcal{O}_{ik}(1)} \tag{21}$$

$$\hat{\boldsymbol{\sigma}}_{ik} = \log\left[\frac{\mathcal{O}_{ik}(1) \cdot \mathcal{O}_{ik}(\boldsymbol{x}^2) - \mathcal{O}_{ik}^2(\boldsymbol{x})}{\mathcal{O}_{ik}^2(1)}\right] \tag{22}$$

$$\hat{\boldsymbol{\omega}}_i = \frac{1}{\sum_{k=1}^{K} \mathcal{O}_{ik}(1)} \cdot [\mathcal{O}_{i1}(1), \ldots, \mathcal{O}_{iK}(1)]'. \tag{23}$$

However, since the general DT objective function $\mathcal{F}(\boldsymbol{\Lambda})$ may be positive definite, negative definite, or even indefinite, the above critical point $\hat{\boldsymbol{\lambda}}$ may be a maximum, a minimum, or a saddle point of $\mathcal{F}(\boldsymbol{\Lambda})$. Even more, it may not exist at all in some special cases. We have conceptually depicted all possible situations in Fig. 1. In total, we can have five different possible cases: 1) $\hat{\boldsymbol{\lambda}}$ is maximum and it is located inside the trust region, as shown in case 1; 2) $\hat{\boldsymbol{\lambda}}$ is maximum but outside the trust region, as in case 2; 3) $\hat{\boldsymbol{\lambda}}$ is a minimum, as in case 3; 4) $\hat{\boldsymbol{\lambda}}$ is a saddle point, as shown in case 4; and 5) no critical point exists, as shown in case 5. Among these cases, even when $\hat{\boldsymbol{\lambda}}$ is indeed a maximum, it may still not be a good solution to (3) since it may be too far from the initial point so that the constraint in (14) is no longer valid, as in case 2.

Our ultimate goal is to optimize the objective function $\mathcal{F}(\boldsymbol{\Lambda})$ subject to the constraints given in (14). We propose to use a line search method to solve the constrained optimization problem. First, we determine a search direction. For cases 1 to 3, we are essentially facing a quadratically constrained quadratic programming (QCQP) [23] problem, and it is intuitive to conduct the line search along the line segment joining the initial point $\boldsymbol{\lambda}^0$ and the calculated critical point $\hat{\boldsymbol{\lambda}}$. However, for cases 4 and 5, it makes more sense to conduct the line search along the gradient direction of the objective function at the initial point $\boldsymbol{\lambda}^0$. In summary, the line search direction $\boldsymbol{d}$ is selected as follows:

$$\boldsymbol{d} = \begin{cases} \hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0 & \text{if } \hat{\boldsymbol{\lambda}} \text{ exists and is not a saddle point} \\ \nabla \mathcal{F}(\boldsymbol{\lambda}^0) & \text{otherwise.} \end{cases} \tag{24}$$

Next, the constrained optimization in (3) is to optimize a scaling coefficient ϵ along the predetermined search direction $\boldsymbol{d}$ as

$$\begin{aligned}&\epsilon^* = \arg\max_{\epsilon} \mathcal{F}[\boldsymbol{\lambda}(\epsilon)] \\ &\text{subject to } \mathcal{D}[\boldsymbol{\lambda}(\epsilon) \,\|\, \boldsymbol{\lambda}^0] \le \rho^2 \end{aligned} \tag{25}$$

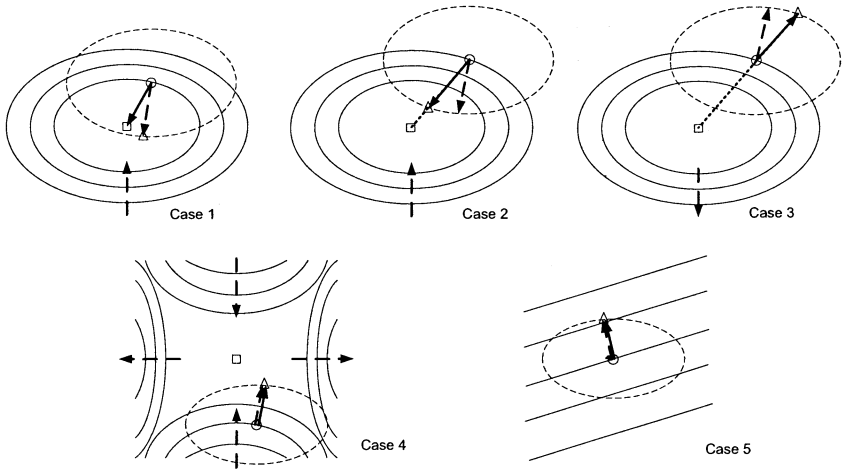


Fig. 1. Illustration of constrained line search for maximizing the objective function in several cases. ($\bigcirc$: $\boldsymbol{\lambda}^0$, the initial point; $\square$: $\hat{\boldsymbol{\lambda}}$, the critical point; $\triangle$: $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}(\epsilon^*)$, the optimal point;—: contours of $\mathcal{F}$; . . .: the trust region; $\leftarrow$: search direction; $\leftarrow$-: gradient direction).

TABLE II
SOLUTIONS FOR CLS WITH QUADRATIC TRUST REGION

case	direction $\boldsymbol{d}$	condition	ϵ^*
1	$\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0$	$\hat{\boldsymbol{\lambda}}$ is a maximum, $Q(\boldsymbol{d},\boldsymbol{\phi}) \le \rho^2$	1
2		$\hat{\boldsymbol{\lambda}}$ is a maximum, $Q(\boldsymbol{d},\boldsymbol{\phi}) > \rho^2$	$+\rho \cdot Q^{-\frac{1}{2}}(\boldsymbol{d},\boldsymbol{\phi})$
3		$\hat{\boldsymbol{\lambda}}$ is a minimum	$-\rho \cdot Q^{-\frac{1}{2}}(\boldsymbol{d},\boldsymbol{\phi})$
4 or 5	$\nabla\mathcal{F}(\boldsymbol{\lambda}^0)$	$\hat{\boldsymbol{\lambda}}$ doesn't exist or it is a saddle point	$+\rho \cdot Q^{-\frac{1}{2}}(\boldsymbol{d},\boldsymbol{\phi})$

TABLE III
CONDITIONS AND CLS FORMULAE TO UPDATE GAUSSIAN MEANS

case	condition	$\boldsymbol{d}$	ϵ^*
1	$\mathcal{O}_{ik}(1) > 0$, $Q(\hat{\boldsymbol{\mu}}_{ik} - \boldsymbol{\mu}^0_{ik}, \boldsymbol{\Sigma}^0_{ik}) < \rho^2$	$\hat{\boldsymbol{\mu}}_{ik} - \boldsymbol{\mu}^0_{ik}$	1
2	$\mathcal{O}_{ik}(1) > 0$, $Q(\hat{\boldsymbol{\mu}}_{ik} - \boldsymbol{\mu}^0_{ik}, \boldsymbol{\Sigma}^0_{ik}) \ge \rho^2$		$+\rho \cdot Q^{-\frac{1}{2}}(\hat{\boldsymbol{\mu}}_{ik} - \boldsymbol{\mu}^0_{ik}, \boldsymbol{\Sigma}^0_{ik})$
3	$\mathcal{O}_{ik}(1) < 0$		$-\rho \cdot Q^{-\frac{1}{2}}(\hat{\boldsymbol{\mu}}_{ik} - \boldsymbol{\mu}^0_{ik}, \boldsymbol{\Sigma}^0_{ik})$
5	$\mathcal{O}_{ik}(1) = 0$	$\nabla\mathcal{F}(\boldsymbol{\mu}^0_{ik})$	$+\rho \cdot Q^{-\frac{1}{2}}(\nabla\mathcal{F}(\boldsymbol{\mu}^0_{ik}), \boldsymbol{\Sigma}^0_{ik})$

where $\boldsymbol{\lambda}(\epsilon) = \boldsymbol{\lambda}^0 + \epsilon \cdot \boldsymbol{d}$ is the model linearly scaled along the direction of $\boldsymbol{d}$.

As long as we impose the quadratic constraints in (14), the above line search problem can be solved efficiently and the optimal scaling coefficient ϵ^* can be computed in a closed-form for all five cases in Fig. 1. For case 1, it is obvious that the optimal coefficient is $\epsilon^* = 1$ since the computed critical point $\hat{\boldsymbol{\lambda}}$ is the solution to (25). For all other cases, it is clear that the optimal point is the intersecting point of the search line with the quadratic constraint surface. In other words, the optimal scaling coefficient ϵ^* satisfies $\mathcal{D}(\boldsymbol{\lambda}^0 + \epsilon^* \cdot \boldsymbol{d} \| \boldsymbol{\lambda}^0) = \rho^2$. After substituting (14) into it, we have

$$\epsilon^{*2} \cdot Q(\boldsymbol{d},\boldsymbol{\phi}) = \rho^2. \tag{26}$$

Therefore, the optimal coefficient ϵ^* can be computed as $\epsilon^* = \pm\rho \cdot Q^{-(1/2)}(\boldsymbol{d},\boldsymbol{\phi})$. Obviously, $\epsilon^* = -\rho \cdot Q^{-(1/2)}(\boldsymbol{d},\boldsymbol{\phi})$ for case 3 while $\epsilon^* = \rho \cdot Q^{-(1/2)}(\boldsymbol{d},\boldsymbol{\phi})$ for cases 2, 4, and 5. The results are summarized in Table II.

It is remarkable that for the problems from cases 1 to 3, the corresponding solutions are intuitive and theoretically grounded according to the convex optimization theories [23], while in the other two nonconvex cases, we are adopting effective first-order algorithm. In the following, based on the above CLS optimization method, we derive the updating formula for Gaussian means, Gaussian variances, and Gaussian mixture weights, respectively.

Because when updating each kernels in one iteration, we come up with a closed form solution, the computational cost of the algorithm is roughly the same compared to EBW.

In some other algorithms for discriminative training such as Quickprop and EBW, gradient and high-order statistics are also used to improved the effectiveness, but heuristic back-off or smoothing mechanisms are also necessary to ensure the reliableness of these statistics. A major difference in our approach is that now the problem is first casted as a constrained optimization, then based on the nature of locality constraint, all the studies can then be conducted in a grounded manner, which provide a new framework to utilize high-order information reliably.

A. Updating Gaussian Means

For Gaussian mean vectors, the critical point $\hat{\boldsymbol{\mu}}_{ik}$ is calculated according to (21). Now we need to examine conditions under which the computed critical point is a maximum, minimum, or saddle point. From (15), it is easy to show that $\nabla^2\mathcal{F}(\boldsymbol{\mu}_{ik}) = \mathcal{O}_{ik}(1) \cdot \boldsymbol{\Sigma}_{ik}^{-1}$. Since $\boldsymbol{\Sigma}_{ik}^{-1}$ is always a positive definite matrix, $\hat{\boldsymbol{\mu}}_{ik}$ cannot be a saddle point. It is a maximum or minimum point depending on the sign of $\mathcal{O}_{ik}(1)$. If $\mathcal{O}_{ik}(1) > 0$, it is a maximum point; otherwise it is a minimum point. If $\mathcal{O}_{ik}(1) = 0$, the objective function $\mathcal{F}(\boldsymbol{\Lambda})$ degenerates into a linear function of $\boldsymbol{\mu}_{ik}$ and the critical point $\hat{\boldsymbol{\mu}}_{ik}$ does not exist.

Furthermore, we can determine whether the computed critical point $\hat{\boldsymbol{\mu}}_{ik}$ satisfies the constraint in (14) by checking $Q(\hat{\boldsymbol{\mu}}_{ik} - \boldsymbol{\mu}^0_{ik}, \boldsymbol{\Sigma}^0_{ik})$: if $Q(\hat{\boldsymbol{\mu}}_{ik} - \boldsymbol{\mu}^0_{ik}, \boldsymbol{\Sigma}^0_{ik}) < \rho^2, \hat{\boldsymbol{\mu}}_{ik}$ locates inside the trust region, as in case 1; otherwise, it locates outside the trust region as in case 2. All these results are summarized in Table III. In each case, the optimal mean vector $\boldsymbol{\mu}^*_{ik}$ is updated as $\boldsymbol{\mu}^*_{ik} = \boldsymbol{\mu}^0_{ik} + \epsilon^* \cdot \boldsymbol{d}$.

TABLE IV
CONDITIONS AND CLS FORMULA TO UPDATE GAUSSIAN VARIANCES

case	condition	$\boldsymbol{d}$	ϵ^*
1	$\mathcal{O}_{ik}(1)\mathcal{O}_{ik}(\boldsymbol{x}^2) > \mathcal{O}^2_{ik}(\boldsymbol{x})$, $Q(\hat{\boldsymbol{\sigma}}_{ik} - \boldsymbol{\sigma}^0_{ik}, \boldsymbol{I}) < \rho^2$	$\hat{\boldsymbol{\sigma}}_{ik} - \boldsymbol{\sigma}^0_{ik}$	1
2	$\mathcal{O}_{ik}(1)\mathcal{O}_{ik}(\boldsymbol{x}^2) > \mathcal{O}^2_{ik}(\boldsymbol{x})$, $Q(\hat{\boldsymbol{\sigma}}_{ik} - \boldsymbol{\sigma}^0_{ik}, \boldsymbol{I}) \geq \rho^2$		$+\rho \cdot Q^{-\frac{1}{2}}(\hat{\boldsymbol{\sigma}}_{ik} - \boldsymbol{\sigma}^0_{ik}, \boldsymbol{I})$
5	$\mathcal{O}_{ik}(1)\mathcal{O}_{ik}(\boldsymbol{x}^2) \leq \mathcal{O}^2_{ik}(\boldsymbol{x})$	$\nabla\mathcal{F}(\boldsymbol{\sigma}^0_{ik})$	$+\rho \cdot Q^{-\frac{1}{2}}(\nabla\mathcal{F}(\boldsymbol{\sigma}^0_{ik}), \boldsymbol{I})$

B. Updating Gaussian Variances

For Gaussian variances, the critical point, $\hat{\boldsymbol{\sigma}}_{ik}$, is calculated according to (22). From (22), we can see that $\hat{\boldsymbol{\sigma}}_{ik}$ exists only when the condition $\mathcal{O}_{ik}(1) \cdot \mathcal{O}_{ik}(\boldsymbol{x}^2) - \mathcal{O}^2_{ik}(\boldsymbol{x}) > 0$ (all the elements in the vector are larger than 0) holds. If $\mathcal{O}_{ik}(1) \cdot \mathcal{O}_{ik}(\boldsymbol{x}^2) - \mathcal{O}^2_{ik}(\boldsymbol{x}) < 0$, we have to conduct line search along gradient direction as in case 5 since the critical point $\hat{\boldsymbol{\sigma}}_{ik}$ does not exist.

Furthermore, based on (16), it is straightforward to show:

$$\nabla^2\mathcal{F}(\boldsymbol{\sigma}_{ik}) = -\frac{1}{2}\mathcal{O}_{ik}(1) \cdot \boldsymbol{\Sigma}^{-1}_{ik} \cdot \exp(\hat{\boldsymbol{\sigma}}_{ik}). \tag{27}$$

If the critical point $\hat{\boldsymbol{\sigma}}_{ik}$ exists, i.e., $\mathcal{O}_{ik}(1) \cdot \mathcal{O}_{ik}(\boldsymbol{x}^2) - \mathcal{O}^2_{ik}(\boldsymbol{x}) > 0$, we can easily derive that $\mathcal{O}_{ik}(1) > 0$. As the result, the second partial derivative w.r.t. $\boldsymbol{\sigma}_{ik}$ in (27) is always negative-definite. Therefore, situations in cases 3 and 4 never happen for Gaussian variances. At last, we summarize all these different conditions and formulae to calculate $\boldsymbol{d}$ and ϵ^* for Gaussian variance in Table IV. Similarly, all variances are updated as $\boldsymbol{\sigma}^*_{ik} = \boldsymbol{\sigma}^0_{ik} + \epsilon^* \cdot \boldsymbol{d}$.

C. Updating Mixture Weights

For Gaussian weights $\boldsymbol{\omega}_i = (\omega_{i1}, \omega_{i2}, \ldots, \omega_{iK})'$, we can obtain the critical point $\hat{\boldsymbol{\omega}}_i$ as in (23), subject to the constraint of $\sum_{k=1}^{K} \omega_{ik} = 1$. It is remarkable that there exist alternative approaches, e.g., to renormalize the weights after free parameter updating, in dealing with the weights. Here we enforce the constraint to keep the updated models are still valid HMMs, and thus a theoretically grounded study can be conducted. Also, it is straightforward to verify that $\hat{\boldsymbol{\omega}}_i$ is a maximum when $\mathcal{O}_{ik}(1) > 0$ for all k, as in case 1 or 2, And $\hat{\boldsymbol{\omega}}_i$ is a minimum when $\mathcal{O}_{ik}(1) < 0$ for all k, as in case 3; otherwise, $\hat{\boldsymbol{\omega}}_i$ is neither maximum nor minimum. In the last case, we follow the gradient to update $\boldsymbol{\omega}_i$. To ensure that the weights remain a valid discrete probability distribution, we must project the gradient in (17) at the initial point, i.e., $\nabla\mathcal{F}(\boldsymbol{\omega}^0_i)$, onto the hyperplane $\sum_{k=1}^{K} \omega_{ik} = 1$, as shown in Fig. 2

$$\nabla\mathcal{F}^{\|}\left(\boldsymbol{\omega}^0_i\right) = \nabla\mathcal{F}\left(\boldsymbol{\omega}^0_i\right) - \left[\nabla\mathcal{F}\left(\boldsymbol{\omega}^0_i\right) \cdot \boldsymbol{u}\right]\boldsymbol{u} \tag{28}$$

where $\boldsymbol{u} = ((1)/(\sqrt{K}), (1)/(\sqrt{K}), \ldots, (1)/(\sqrt{K}))'$ is the normal vector of the hyperplane. We summarized all these different conditions and formula to calculate $\boldsymbol{d}$ and ϵ^* for Gaussian weights in Table V. Similarly, weights is updated as $\boldsymbol{\omega}^*_{ik} = \boldsymbol{\omega}^0_{ik} + \epsilon^* \cdot \boldsymbol{d}$.

In practice, we should also check the boundary condition of $0 < \omega_{ik} < 1, 1 \leq k \leq K$ to ensure a valid discrete probability distribution.

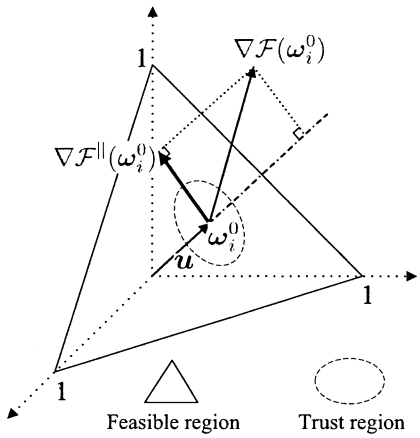


Fig. 2. Illustration of solving CLS problems for weight vectors by using the projected gradient decent.

TABLE V
CONDITIONS AND CLS FORMULA TO UPDATE GAUSSIAN WEIGHTS

case	condition	$\boldsymbol{d}$	ϵ^*
1	$\mathcal{O}_{ik}(1) > 0$, $Q(\hat{\boldsymbol{\omega}}_i - \boldsymbol{\omega}^0_i, \boldsymbol{\Pi}^0_i) < \rho^2$	$\hat{\boldsymbol{\omega}}_{ik} - \boldsymbol{\omega}^0_{ik}$	1
2	$\mathcal{O}_{ik}(1) > 0$, $Q(\hat{\boldsymbol{\omega}}_i - \boldsymbol{\omega}^0_i, \boldsymbol{\Pi}^0_i) \geq \rho^2$		$+\rho \cdot Q^{-\frac{1}{2}}(\hat{\boldsymbol{\omega}}_i - \boldsymbol{\omega}^0_i, \boldsymbol{\Pi}^0_i)$
3	$\mathcal{O}_{ik}(1) < 0$		$+\rho \cdot Q^{-\frac{1}{2}}(\hat{\boldsymbol{\omega}}_i - \boldsymbol{\omega}^0_i, \boldsymbol{\Pi}^0_i)$
4	otherwise	$\nabla\mathcal{F}^{\|}(\boldsymbol{\omega}^0_i)$	$+\rho \cdot Q^{-\frac{1}{2}}(\nabla\mathcal{F}^{\|}(\boldsymbol{\omega}^0_i), \boldsymbol{\Pi}^0_i)$

V. EXPERIMENTS

In order to verify the effectiveness of the proposed CLS optimization method, we evaluated it on several benchmark speech recognition tasks, including: connected digit string recognition with the TIDIGITS database, continuous speech recognition with the Resource Management (RM) database, and large-vocabulary continuous speech recognition with the Switchboard database (both mini-train set and full *h5train00* set [1]). Experimental setups are summarized in Table VI.

Because EBW is the most popular and successful method for optimizing the MMI and other DT criteria, such as MPE and MD, in our experiments, the CLS method is only compared with it. In our EBW implementation, following [8], we use kernel dependent smoothing factors which are set to be twice of the corresponding denominator occupancy. When we use EBW for the MPE training, we also use I-smoothing [8] with τ setting to 100 during each iteration. In CLS experiments, the most recently obtained models are set as the initial model set $\boldsymbol{\Lambda}^0$, and the model parameters are updated according to the CLS formula in Section IV.

TABLE VI
EXPERIMENTAL SETUP IS LISTED FOR ALL RECOGNITION TASKS

Training set		Acoustic features	# tied states	# kernels/state
TIDIGITS		13 MFCCs $+\Delta+\Delta\Delta$	114	6
Switch-board	*mini-train*	13 PLPs $+\Delta+\Delta\Delta$	1500	12
	h5train00	13 PLPs $+\Delta+\Delta\Delta$	6000	16
RM		13 MFCCs $+\Delta+\Delta\Delta$	1600	6

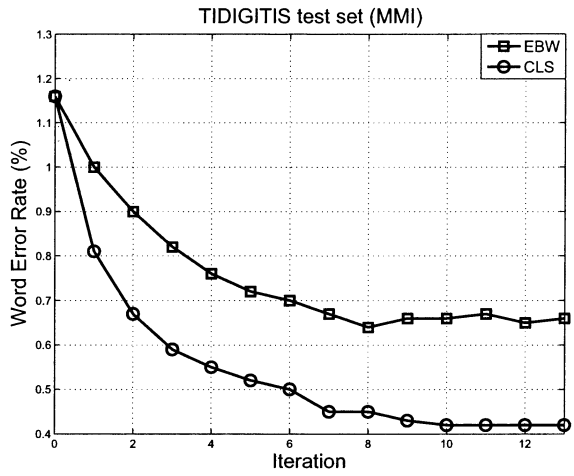


Fig. 3. Comparison of word error rates of different optimization methods in the MMI training on the TIDIGITS test set.

The constant ρ^2 is set to $0.1/n$ for nth iteration for all the parameters. Actually, the physical meaning of KLD constraints ensures that it is statistically normalized, and the same trust region control parameter can be used across models. Hence, we just verified that the scheme of selecting ρ^2 introduced above leads to smooth updating procedure on TIDIGITS and then adopt it for all the experiments.

A. TIDIGITS

The TIDIGITS database contains utterances from a total of 326 speakers (111 men, 114 women, and 101 children). In our experiments, we used all data from adults and children, which includes totally 12 549 training utterances and 12 547 testing utterances. The vocabulary is composed of 11 digits of "zero" to "nine," and "oh." The length of digit strings varies from one to seven digits. Each digit is modeled by a ten-state, left-to-right, whole-word Gaussian mixture CDHMMs. The baseline ML model consists of 114 tied states with six Gaussians per state. The acoustic feature and model size are summarized in Table VI. In the experiment, the ML model is used as the seed model for discriminative training, in either EBW or CLS method.

In Fig. 3, we compare the learning curves of CLS and EBW methods in the MMI training. The results clearly show that the proposed CLS method yields better performance than the EBW method. The CLS algorithm shows both a faster convergence and a lower recognition error rate than the conventional EBW algorithm. For CLS, word error rate decreases from 1.16% of the ML baseline performance to 0.42%, or a 63.8% relative error reduction. Meanwhile, the EBW algorithm only achieves 44% relative error reduction.

TABLE VII
SUMMARY OF RECOGNITION PERFORMANCE IN TIDIGITS BY USING EBW OR CLS OPTIMIZATION METHOD FOR MMI AND MD TRAINING CRITERIA

Criterion	Optimization	WER (in %)
ML	BW	1.16
MMI	EBW	0.65
	CLS	0.42
MD	EBW	0.44
	CLS	0.40

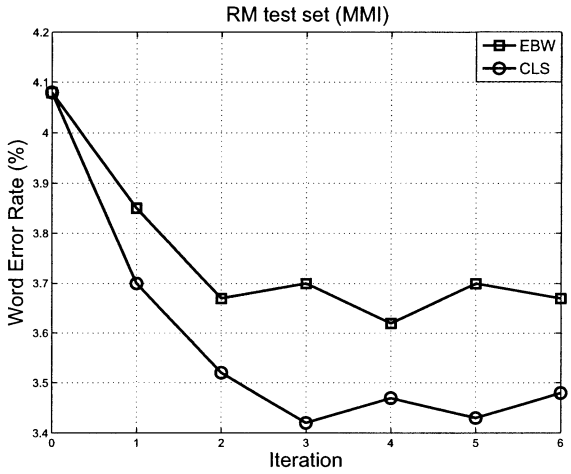


Fig. 4. Comparison of word error rates (in %) of different optimization methods on the Resource Management test set, based on MMI criterion.

In addition, we compare CLS with EBW for the MD criterion, and the results are shown in Table VII. Again, CLS outperforms EBW. In the MD training, the CLS algorithm achieves a 0.40% word error rate which is slightly better than 0.44% obtained by the EBW method.

B. Resource Management(RM)

In this section, we examine the CLS algorithm in a medium size continuous speech recognition task on the DARPA Resource Management (RM) database. In this experiment, we only use the speaker independent portion in the training data set. And the test data is composed of the data sets of Feb'89, Oct'89, Feb'92, and Sep'92, with 300 sentences in each set. Context-dependent triphone CDHMMs are used. The best ML-trained models set has 1600 tied states with six Gaussians per state. Acoustic features and model size are summarized in Table VI as well. The ML model is used as the initial model in discriminative training. Word-pair language model is adopted for testing. In order to obtain richer acoustically competing hypotheses, we use unigram language models to decode training data in word lattice generation.

In Fig. 4, we compare the learning curves of the EBW and CLS algorithms in MMI training. The results show that CLS outperforms EBW where CLS decreases the word error rate from 4.08% to 3.43%, with a 16.2% relative error reduction from the ML baseline.

Besides the MMI training, we also compare CLS with EBW for optimizing the MPE criterion. The results are shown in

TABLE VIII
SUMMARY OF RECOGNITION PERFORMANCE IN RM BY USING EBW OR CLS OPTIMIZATION METHOD FOR MMI AND MPE TRAINING CRITERIA

Criterion	Optimization	WER (in %)
ML	BW	4.08
MMI	EBW	3.62
	CLS	3.43
MPE	EBW w/ I-smooting	3.87
	CLS	3.39

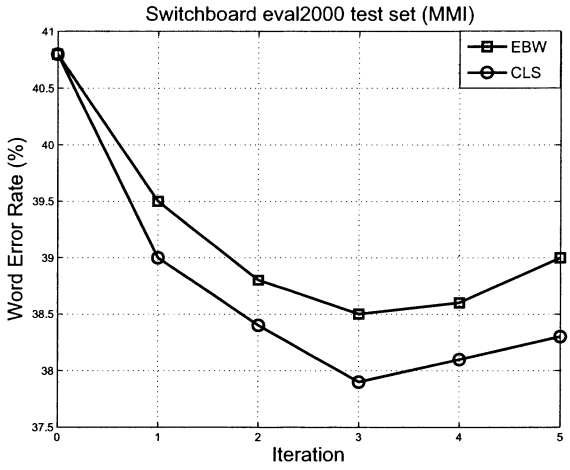


Fig. 5. Comparison of word error rates of different optimization methods on the Switchboard *eval2000* test set, using *mini-train* training set, based on MMI criterion.

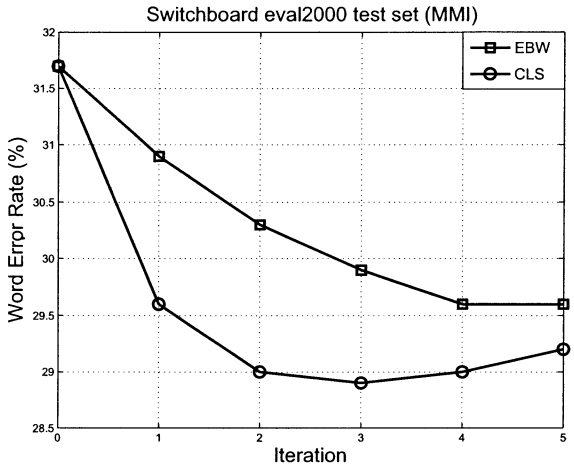


Fig. 6. Comparison of word error rates of different optimization methods on the Switchboard *eval2000* test set, using *h5train00* training set, based on MMI criterion.

TABLE IX
SUMMARY OF RECOGNITION PERFORMANCE (WER IN %) IN SWITCHBOARD BY USING EBW OR CLS OPTIMIZATION METHOD FOR MMI AND MPE TRAINING CRITERIA

Criterion	Optimization	mini-train	full *h5train00*
ML	BW	40.8	31.7
MMI	EBW	38.5	29.6
	CLS	37.9	28.9
MPE	EBW w/ I-smooting	38.0	28.7
	CLS	37.7	28.4

Table VIII. In the MPE training, we still observe that the CLS algorithm achieves bigger improvement than EBW and CLS achieves 3.39% word error rate while EBW obtains 3.87% word error rate.

C. Switchboard

In the Switchboard recognition task, we used two different training sets: the *mini-train* and the full *h5train00* set, consisting of 18 and 265 h of speech data, respectively. Both training sets contain data from the Switchboard I (SWB1) corpus and the *h5train00* set also contains Call Home English (CHE) data. *Eval2000* set, which contains 1831 utterances, has been used as the evaluation set. Context-dependent triphone HMMs are used in the following experiments. Trigram language model is used in evaluation but unigram language model is used in training to generate word lattices for discriminative training. The NIST scoring software [21] has been used to evaluate word error rates. The baseline ML models are estimated using the standard Baum–Welch (BW) algorithm. The experimental setup, including acoustic features and the best ML model size, is summarized in Table VI.

We first compare CLS with EBW in MMI training. The learning curves are shown in Figs. 5 and 6 for the *mini-train* and *h5train00* sets. Once again, the results show that the proposed CLS algorithm achieves better word accuracy and faster convergence than the EBW method on both *mini-train* and *h5train00* sets. Comparing with the ML baseline, the CLS training algorithm reduces word error rate from 40.8% to 37.9%, or a 7.1% relative error reduction for the *mini-train* set, and from 31.7% to 28.9%, or a 8.8% relative error reduction for the *h5train00 set*, respectively. Comparing CLS with EBW in MMI training, we observe that CLS outperforms EBW on both *mini-train* and *h5train00* sets, where the EBW only achieves 38.5% word error rate in *mini-train* and 29.6% word error rate in *h5train00*.

Finally, we also compare CLS with EBW for MPE training. The results are summarized in Table IX. It is clearly shown that CLS yields better recognition performance than EBW in the MPE training as well. For example, the CLS training achieves 37.7% word error rate for *mini-train* and 28.4% word error rate for *h5train00* while the EBW method obtains only 38.0% word error rate for *mini-train* and 28.7% for full *h5train00* set.

VI. CONCLUSION

In this paper, a new optimization method, called CLS, is proposed for discriminative training of speech recognition HMMs. The proposed CLS method is general enough to optimize various popular objective functions in discriminative training. In this paper, discriminative training of CDHMMs is first formulated as a constrained optimization problem, where a constraint is imposed based on the KLD between models, which guarantees an equalized updating process across all the parameters in the model set. Based upon some approximations on the

KLD constraint, closed-form solutions can be easily derived for updating various CDHMM parameters. We examined the proposed CLS methods on several standard speech recognition tasks, from small-vocabulary digit string recognition to large-vocabulary continuous speech recognition. Experimental results clearly show that the proposed CLS method consistently yields better performance than the popular EBW method for all examined discriminative training criteria.

Although the constrained optimization framework provide us a meaningful framework in solving the stableness issue of discriminative training, we will further study various practical issues under this framework. For example, how to set theoretically grounded trust regions across models, and how to update variances more reliably.

APPENDIX

In this Appendix, we calculate partial derivatives of the general DT objective function $\mathcal{F}(\mathbf{\Lambda})$ with respect to any CDHMM parameter. The general DT objection function can be expressed, as shown in (29) at the bottom of the page.

For any model parameter, denoted as $\boldsymbol{\lambda}_{ik}$[1], we have

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\lambda}_{ik}}\mathcal{F}(\mathbf{\Lambda}) &= \frac{1}{R}\sum_{r=1}^{R} f_r' \cdot \frac{1}{\kappa} \cdot \sum_{W\in\mathcal{M}_r} \\ &\times \Bigg[\underbrace{\frac{p^{\kappa}(\boldsymbol{X}_r \mid W)p(W)G(W,W_r)}{\sum_{W'\in\mathcal{M}_r} p^{\kappa}(\boldsymbol{X}_r \mid W')p(W')G(W',W_r)}}_{G(W,W_r \mid \boldsymbol{X}_r)} \\ &\times \frac{\partial \log p(\boldsymbol{X}_r \mid W)}{\partial \boldsymbol{\lambda}_{ik}} \\ &- \underbrace{\frac{p^{\kappa}(\boldsymbol{X}_r \mid W)p(W)}{\sum_{W'\in\mathcal{M}_r} p^{\kappa}(\boldsymbol{X}_r \mid W')p(W')}}_{p(W \mid \boldsymbol{X}_r)} \frac{\partial \log p(\boldsymbol{X}_r \mid W)}{\partial \boldsymbol{\lambda}_{ik}} \Bigg]\end{aligned} \tag{30}$$

where we denote

$$f_r' = f'\left(\log\left[\frac{\sum_{W\in\mathcal{M}_r} p^{\kappa}(\boldsymbol{X}_r \mid W_r)p(W_r)G(W,W_r)}{\sum_{W'\in\mathcal{M}_r} p^{\kappa}(\boldsymbol{X}_r \mid W')p(W')}\right]^{\frac{1}{\kappa}}\right). \tag{31}$$

[1]Here $\boldsymbol{\lambda}_{ik}$ represents the ith state, kth Gaussian component of Gaussian mean vector, covariance matrix, or mixture weight

When the smoothing factor κ is sufficiently small ($\kappa \to 0$) and the models do not deviate too much from the values in each iteration, it is reasonable to assume that all the three terms, i.e., f_r', $G(W, W_r \mid \boldsymbol{X}_r)$, and $p(W \mid \boldsymbol{X}_r)$, are approximately constants with respect to model parameter $\boldsymbol{\lambda}_{ik}$ since they are all related to model parameters $\boldsymbol{\lambda}_{ik}$ only through $p^{\kappa}(\boldsymbol{X}_r \mid W_r)$. Accordingly, we have

$$\begin{aligned}\nabla\mathcal{F}(\lambda_{ik}) &\equiv \frac{\partial \mathcal{F}(\mathbf{\Lambda})}{\partial \boldsymbol{\lambda}_{ik}} \\ &= \frac{1}{R}\sum_{r=1}^{R}\sum_{W\in\mathcal{M}_r} C_r(W)\frac{\partial \log p(\boldsymbol{X}_r \mid W)}{\partial \boldsymbol{\lambda}_{ik}}\end{aligned} \tag{32}$$

where we denote $C_r(W) = (f_r')/(\kappa) \cdot [G(W, W_r \mid \boldsymbol{X}_r) - p(W \mid \boldsymbol{X}_r)]$. Furthermore, we have

$$\frac{\partial \log p(\boldsymbol{X}_r \mid W)}{\partial \boldsymbol{\lambda}_{ik}} = \sum_{t=1}^{T}\gamma_{ik}^{W}(r,t)\frac{\partial \log[\omega_{ik}\mathcal{N}(\boldsymbol{x}_{rt};\boldsymbol{\mu}_{ik},\mathbf{\Sigma}_{ik})]}{\partial \boldsymbol{\lambda}_{ik}} \tag{33}$$

where $\gamma_{ik}^{W}(r,t)$ denotes posterior probabilities collected for kth Gaussian component in ith state of the composite HMM corresponding to W based on $\boldsymbol{X}_r$. After substituting (33) into (32), we obtain the following formula:

$$\begin{aligned}\nabla\mathcal{F}(\boldsymbol{\lambda}_{ik}) = \frac{1}{R}\sum_{r=1}^{R}\sum_{W\in\mathcal{M}_r} C_r(W)\sum_{t=1}^{T}\gamma_{ik}^{W}(r,t) \\ \cdot\frac{\partial \log[\omega_{ik}\mathcal{N}(\boldsymbol{x}_{rt} \mid \boldsymbol{\mu}_{ik},\mathbf{\Sigma}_{ik})]}{\partial \boldsymbol{\lambda}_{ik}}\end{aligned} \tag{34}$$

REFERENCES

[1] 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.

[2] 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.

[3] H. Jiang, F. Soong, and C.-H. Lee, "A dynamic in-search data selection method with its applications to acoustic modeling and utterance verification," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 945–955, Sep. 2005.

[4] B. Liu, H. Jiang, J.-L. Zhou, and R.-H. Wang, "Discriminative training based on the criterion of least phone competing tokens for large vocabulary speech recognition," in *Proc. ICASSP'05*, Philadelphia, PA, 2005, pp. 117–120.

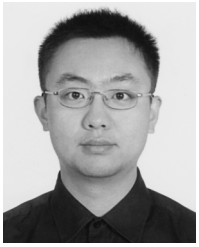
[5] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP'02*, Orlando, FL, 2002, pp. 105–108.

[6] J. Du, P. Liu, F. K. Soong, J.-L. Zhou, and R.-H. Wang, "Minimum divergence based discriminative training," in *Proc. ICSLP'06*, 2006, pp. 2410–2413.

[7] Y. Normandin and H. M. Models, "Maximum mutual information estimation, and the speech recognition problem," Ph.D. dissertation, Dept. Elect. Eng., McGill Univ., Montreal, QC, Canada, 1991.

$$\mathcal{F}(\mathbf{\Lambda}) = \frac{1}{R}\cdot\sum_{r=1}^{R} f\left(\log\left[\frac{\sum_{W\in\mathcal{M}_r} p^{\kappa}(\boldsymbol{X}_r \mid W)p(W)G(W,W_r)}{\sum_{W'\in\mathcal{M}_r} p^{\kappa}(\boldsymbol{X}_r \mid W')p(W')}\right]^{\frac{1}{\kappa}}\right) \tag{29}$$

[8] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Cambridge Univ., Cambridge, U.K., 2004.
[9] X. He and W. Chou, "Minimum classification error linear regression for acoustic model adaptation of continuous density HMMs," in *Proc. ICASSP'03*, 2003, pp. 556–559.
[10] E. McDermott, T. J. Hazen, J. Le 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.
[11] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, pp. 79–86, 1951.
[12] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "Generalization of the Baum algorithm to rational objective functions," in *Proc. ICASSP'89*, 1989, pp. 631–634.
[13] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP'84*, San Diego, CA, 1984, pp. 42.11.1–42.11.4.
[14] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," in *Proc. ICASSP'88*, 1988, pp. 651–654.
[15] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. ICASSP'92*, 1992, pp. 517–520.
[16] R. Schluter, "Investigations on discriminative training criteria," Ph.D. dissertation, Aachen Univ., Aachen, Germany, 2000.
[17] M. Afify, X. W. Li, and H. Jiang, "Statistical performance analysis of MCE/GPD learning in Gaussian classifiers and hidden Markov models," in *Proc. ICASSP'05*, Philadelphia, PA, 2005, pp. 105–108.
[18] M. Afify, X.-W. Li, and H. Jiang, "Statistical analysis of minimum classification error learning for Gaussian and hidden Markov model classifiers," *Trans. Audio, Speech, Lang. Process.*, 15, no. 8, pp. 2405–2417, Nov. 2007.
[19] L. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occuring in the statistical analysis of probablistic functions of markov chains," *Ann. Math. Statist.*, vol. 41, no. 1, pp. 164–171, 1970.
[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[21] D. S. Pallett, W. M. Fisher, and J. G. Fiscus, "Tools for the analysis of benchmark speech recognition tests," in *Proc. ICASSP'90*, 1990, pp. 97–100.
[22] F. J. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 288–298, Mar. 2001.
[23] H. Hindi, "A tutorial on convex optimization," in *Proc. Amer. Control Conf.*, Boston, MA, Jun. 2004, pp. 2352–2365.



Peng Liu received the Ph.D. degree from Tsinghua University, Beijing, China, in 2005.

Since March 2005, he has been an Associate Researcher with Microsoft Research Asia, Beijing. His current research interests include speech and audio processing, handwriting recognition, statistical modeling, and multimodal user interfaces, especially in discriminative training and acoustic modeling.

Cong Liu received the B.A. degree from the University of Science and Technology of China, Hefei, in 2005, where he is currently pursuing the Ph.D. degree in electronic information engineering.

Since September 2004, he has been with the iFlytek Speech Lab, University of Science and Technology of China. From July 2006 to December 2006, he was a visiting student at Microsoft Research Asia, Beijing. His research interests include confidence measures, discriminative training in speech recognition, and also speaker recognition.

Hui Jiang (M'00) received the B.Eng. and M.Eng. degrees from the University of Science and Technology of China, Hefei, 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 worked as a Researcher with the University of Tokyo. From April 1999 to June 2000, he was with Department of Electrical and Computer Engineering, University of Waterloo, Canada as a postdoctoral fellow. From 2000 to 2002, he worked with 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 the fall of 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.

Frank K. Soong (SM'91) received the B.S. degree from National Taiwan University, Taipei, Taiwan, R.O.C., the M.S. degree from the University of Rhode Island, Kingston, and the Ph.D. degree from Stanford University, Stanford, CA, all in electrical engineering.

He joined Bell Labs Research, Murray Hill, NJ, in 1982, and worked there for 20 years and retired as a Distinguished Member of Technical Staff in 2001. With Bell Labs, he had worked on various aspects of acoustics and speech, processing, including speech coding, speech and speaker recognition, stochastic modeling of speech signals, efficient search algorithms, discriminative training, dereverberation of audio and speech signals, microphone array processing, acoustic echo cancellation, and hands-free noisy speech recognition. He was also responsible for transferring recognition technology from research to AT&T voice-activated cell phones which were rated by the Mobile Office Magazine as the best among competing products evaluated. He visited Japan twice as a Visiting Researcher, first from 1987 to 1988, at the NTT Electro-Communication Labs, Musashino, Tokyo, then, from 2002–2004 at the Spoken Language Translation Labs, ATR, Kyoto. In 2004, he joined Microsoft Research Asia (MSRA), Beijing, China, to lead the Speech Research Group. He is a Visiting Professor of the Chinese University of Hong Kong (CUHK) and the codirector of CUHK-MSRA Joint Research Lab, recently promoted to a National Key Lab of Ministry of Education, China. He has published extensively and coauthored more than 150 technical papers in the speech and signal processing fields.

Dr. Soong was the corecipient of the Bell Labs President Gold Award for developing the Bell Labs Automatic Speech Recognition (BLASR) software package. He was the cochair of the 1991 IEEE International Arden House Speech Recognition Workshop. He has served in the IEEE Speech and Language Processing Technical Committee of the Signal Processing Society, as a committee member and Associate Editor of the TRANSACTIONS OF SPEECH AND AUDIO PROCESSING.

Ren-Hua Wang (M'08) was born in Shanghai, China, in August 1943.

He is currently a Professor and Ph.D. supervisor in the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei. His research interests include speech coding, speech synthesis and recognition, and multimedia communication. During the past 20 years, he was in charge of more than ten national key research projects in the information field.

Dr. Wang is a recipient of the 2002 Second Class National Award for Science and Technology Progress, China, and 2005 Information Industries Significant Technology Award for Invention, China.