A CONVEX OPTIMIZATION METHOD FOR JOINT MEAN AND VARIANCE PARAMETER ESTIMATION OF LARGE-MARGIN CDHMM

Tsung-Hui Chang[†]*, Zhi-Quan Luo*[‡]*, Li Deng*[⋆]*, and Chong-Yung Chi*[†]

[†]Institute of Commun. Eng.
National Tsing Hua University,
Hsinchu, Taiwan 30013, R.O.C.
E-mail: d915691@oz.nthu.edu.tw
cychi@ee.nthu.edu.tw

[‡]Dept. of Elect. & Comp. Eng.
University of Minnesota,
Minneapolis, MN 55455, U.S.A.
E-mail: luozq@ece.umn.edu

[⋆]Microsoft Research
Microsoft Corporation,
Redmond, WA 98052, U.S.A.
E-mail: deng@microsoft.com

ABSTRACT

In this paper, we develop a new class of parameter estimation techniques for the Gaussian Continuous-Density Hidden Markov Model (CDHMM), where the discriminative margin among a set of HMMs is used as the objective function for optimization. In addition to optimizing the mean parameters of the large-margin CDHMM, which was attempted in the past, our new technique is able to optimize the variance parameters as well. We show that the joint mean and variance estimation problem is a difficult optimization problem but can be approximated by a convex relaxation method. We provide some simulation results using synthetic data which possess key properties of speech signals to validate the effectiveness of the new method. In particular, we show that with joint optimization of the mean and variance parameters, the CDHMMs under model mismatch are much more discriminative than with only the mean parameters.

Index Terms— Classification, Gaussian CDHMM, Large margin parameter estimation, Convex optimization

1. INTRODUCTION

Parameter estimation of the Gaussian Continuous-Density Hidden Markov Model (CDHMM) [1] is one of the important techniques in automatic speech recognition (ASR), where the speech signals are modeled by CDHMMs. Conventionally, the parameters of the HMM are learnt from the training data using the maximum likelihood estimator (MLE) [1]. The MLE does not minimize recognition error rates because HMMs themselves are inaccurate speech signal models, and often the amount of training data is not very sufficient. In order to overcome the weaknesses of the MLE, ASR researchers have proposed the use of discriminative training criteria (e.g., [2, 3]). The training methods based on these criteria, however, have been aimed to find classification boundaries that minimize empirical error rates on training sets, but they may not generalize well to unseen test sets due to the mismatch between CDHMMs and real speech signals.

It was shown in the seminal work of [4] that the test-set error rate is closely related to the margins (i.e., the distances between the well classified samples and the decision boundary), and the model parameters which possess larger values of margins can exhibit better generalization ability. The concept of large margin has been successfully used in designing state-of-the-art multi-way classifiers for many years [4, 5]. Its application to estimating the CDHMM parameters for ASR, however, is more recent [6–9]. One specific incorporation of the margin [6–9] into CDHMM parameter estimation, referred to as the *large-margin* CDHMM, was shown to be capable of reducing ASR error rates compared with the MLE and margin-free parameter estimation methods.

Since the large-margin CDHMM as proposed in [7–9] estimates only the means in the normalized Gaussian distribution of CDHMM, in the paper we consider a new technique for joint estimation of mean and variance parameters. The importance of the variance parameters in CDHMM for the performance of ASR has been well known. For the MLE, it is straightforward to estimate the variance parameters [1], but for the objective function of the large-margin CDHMM, the estimation problem is much more difficult. Specifically, the associated problem can be shown to be a nonconvex optimization problem with an indefinite quadratic objective function [10]. The proposed method is based on careful reformulation of the associated optimization problem followed by an approximation technique of convex relaxation, which is presented in Section 3. Thus an approximate solution can be efficiently obtained by solving a convex optimization problem through modern optimization algorithms [11]. In Section 4, some simulation results are presented to validate the effectiveness of the presented estimation technique.

2. PROBLEM STATEMENT

Consider an ASR task where there are M words $\{W_1, \ldots, W_M\}$. Given the signal parameter set $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_m\}_{m=1}^M$ where $\boldsymbol{\theta}_m$ is the parameter set of word W_m, the goal of ASR is to classify the observed utterance into one of the words according to the maximum a posteriori criteria [1]. Let $\boldsymbol{X}_{\rm tr} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_{N_T}\}$ be a sequence of training utterances, where $\boldsymbol{X}_k = \{\boldsymbol{x}_{k,1}, \ldots, \boldsymbol{x}_{k,T}\}$, $\boldsymbol{x}_{k,t} \in \mathbb{R}^D$ is a $D \times 1$ vector measurement [2], T is the utterance length, and N_T is the number of training utterances. Let $P(\boldsymbol{X}_k|\boldsymbol{\theta}_m)$ be the conditional probability density function (pdf) of $\boldsymbol{X}_k$ given $\boldsymbol{\theta}_m$, and $P(W_m)$ be the occurrence probability of word W_m. Suppose that $\boldsymbol{X}_k$ is of word W_{i_k}, $i_k \in \Omega \triangleq \{1, \ldots, M\}$. The observation $\boldsymbol{X}_k$ will be correctly classified into W_{i_k} if and only if

$$g_{i_k}(\boldsymbol{X}_k, \boldsymbol{\Theta}) \triangleq \mathcal{F}(\boldsymbol{X}_k|\boldsymbol{\theta}_{i_k}) - \max_{m\in\Omega, m\neq i_k} \mathcal{F}(\boldsymbol{X}_k|\boldsymbol{\theta}_m) \geq 0, \quad (1)$$

where $\mathcal{F}(\boldsymbol{X}_k|\boldsymbol{\theta}_m) = \log(P(W_m)) + \log(P(\boldsymbol{X}_k|\boldsymbol{\theta}_m))$ is the discriminative function [8]. The function $g_{i_k}(\boldsymbol{X}_k, \boldsymbol{\Theta})$ is called the *margin* of $\boldsymbol{X}_k$ associated with the model $\boldsymbol{\Theta}$. Recent advances in statistical learning theory [4, 5] have revealed that the $\boldsymbol{\Theta}$ which corresponds

[†]This work is supported by National Science Council, R.O.C., under Grants NSC 96-2628-E-007-002-MY2 and NSC 96-2219-E-007-001.

[‡]This work is supported in part by the U.S. NSF under Grant DMS-0610037 and in part by the USDOD ARMY under Grant W911NF-05-1-0567.

to larger values of $g_{i_k}(\boldsymbol{X}_k, \boldsymbol{\Theta})$ can exhibit better generalization capability. The model parameters which own this large margin property is called the *large-margin CDHMM* [8]; while the associated estimator is called the large margin estimator (LME).

The large-margin CDHMM can be obtained by maximizing the minimum positive value of $g_{i_k}(\boldsymbol{X}_k, \boldsymbol{\Theta})$ [7–9]. Typically, we only need consider those utterances which are relatively "close" to the decision boundary. Define the index subset

$$\mathcal{S} = \{k \mid 0 \le g_{i_k}(\boldsymbol{X}_k, \boldsymbol{\Theta}) \le \epsilon,\ k \in \{1, \ldots, N_T\}\}, \qquad (2)$$

where $\epsilon > 0$ is a preset number. The collection of $\{\boldsymbol{X}_k | k \in \mathcal{S}\}$ is called the support token set [8]. Mathematically, the LME can be formulated as the following optimization problem [7–9]

$$\begin{aligned}\boldsymbol{\Theta}_{\mathrm{LME}} &= \arg\max_{\boldsymbol{\Theta}} \min_{k\in\mathcal{S}} \left\{ \mathcal{F}(\boldsymbol{X}_k|\boldsymbol{\theta}_{i_k}) - \max_{m\in\Omega, m\neq i_k} \mathcal{F}(\boldsymbol{X}_k|\boldsymbol{\theta}_m) \right\} \\ &= \arg\min_{\boldsymbol{\Theta}} \max_{\substack{m\in\Omega, m\neq i_k \\ k\in\mathcal{S}}} \left\{ \mathcal{F}(\boldsymbol{X}_k|\boldsymbol{\theta}_m) - \mathcal{F}(\boldsymbol{X}_k|\boldsymbol{\theta}_{i_k}) \right\}. \qquad (3)\end{aligned}$$

In ASR, the utterance $\boldsymbol{X}_k = \{\boldsymbol{x}_{k,1}, \ldots, \boldsymbol{x}_{k,T}\}$ is widely modeled by Gaussian CDHMM [1]. Assume that each state in the HMM is modeled by a (single-mixture) multivariate Gaussian random variable. Denote by N the total number of states in the HMM, and let $\{k_1(m), \ldots, k_T(m)\}$, $k_t(m) \in \{1, \ldots, N\}$, be the Viterbi state sequence in $P(\boldsymbol{X}_k|\boldsymbol{\theta}_m)$ [1]. The discriminative function $\mathcal{F}(\boldsymbol{X}_k|\boldsymbol{\theta}_m)$ can be approximated as [1, 8]

$$\begin{aligned}\mathcal{F}(\boldsymbol{X}_k|\boldsymbol{\theta}_m) \approx \gamma_k(m) &- \sum_{t=1}^{T}\sum_{d=1}^{D} \log(\sigma_{k_t(m),d}) \\ &- \sum_{t=1}^{T}\sum_{d=1}^{D} \frac{1}{2}\left(\frac{x_{k,t,d} - \mu_{k_t(m),d}}{\sigma_{k_t(m),d}}\right)^2, \qquad (4)\end{aligned}$$

where $\{\mu_{k_t(m),d}, \sigma^2_{k_t(m),d}\}_{d=1}^D$ are[1] the means and variances of state $k_t(m)$ associated with word W_m, and $\gamma_k(m) = \log(\pi_{k_1(m)}) + \sum_{t=2}^{T}\log(a_{k_{t-1}(m),k_t(m)}) - \frac{TD}{2}\log(2\pi) + \log(P(W_m))$, in which $\{\pi_{k_t(m)}\}$ and $\{a_{k_{t-1}(m),k_t(m)}\}$ are the initial and transition probabilities of word W_m, respectively.

From (3) and (4), one can see that the LME problem is a difficult optimization problem. Therefore, most of the existing methods [7–9] focus only on the mean parameter optimization. Specifically, it has been shown [7, 9] that, if only the mean parameters are considered, the LME problem in (3) can be approximated by a semidefinite program (SDP). It is intuitive to obtain a better margin distribution by jointly optimizing the mean and variance parameters. However, the associated problem becomes much more difficult. In the next section a new estimation technique is presented for achieving this goal. In the paper, we refer to the LME which only optimizes the mean parameters as the large margin mean estimator (LMME); while the one which optimizes both mean and variance parameters as the large margin mean and variance estimator (LMMVE).

3. LARGE MARGIN MEAN AND VARIANCE ESTIMATION

Let us redefine $\boldsymbol{\theta}_m = \{\{\mu_{m,n,d}, \sigma^2_{m,n,d}\}_{d=1}^D, n = 1, \ldots, N\}$ as the set containing the parameters of interest of W_m. According to [7, 8], the LME problem (3) would be unbounded below if there is no proper constraints on $\boldsymbol{\Theta}$. To fix this, we introduce two spherical constraints

$$\sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{d=1}^{D}\left(\frac{\mu_{m,n,d} - \bar{\mu}_{m,n,d}}{\sigma_{m,n,d}}\right)^2 \le r^2, \qquad (5a)$$

$$|\sigma_{m,n,d} - \bar{\sigma}_{m,n,d}| \le \beta, \qquad (5b)$$

$$m = 1, \ldots, M,\ n = 1, \ldots, N,\ d = 1, \ldots, D,$$

where $\{\bar{\mu}_{m,n,d}, \bar{\sigma}^2_{m,n,d}\}$ serve as the "sphere center" and can be the estimates of the MLE or any other discriminative training estimators. Equation (5) means that the large-margin CDHMM is obtained by searching inside the sphere with "radius" $r \ge 0$ for the mean parameters, and with radius $\beta \ge 0$ for the variance parameters. By (5), we can rewrite the LME problem (3) as the following constrained optimization problem

$$\min_{\boldsymbol{\Theta}} \max_{\substack{m\in\Omega, m\neq i_k \\ k\in\mathcal{S}}} \left\{ \mathcal{F}(\boldsymbol{X}_k|\boldsymbol{\theta}_m) - \mathcal{F}(\boldsymbol{X}_k|\boldsymbol{\theta}_{i_k}) \right\} \qquad (6a)$$

$$\text{subject to (s.t.)} \sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{d=1}^{D}\left(\frac{\mu_{m,n,d} - \bar{\mu}_{m,n,d}}{\sigma_{m,n,d}}\right)^2 \le r^2 \qquad (6b)$$

$$|\sigma_{m,n,d} - \bar{\sigma}_{m,n,d}| \le \beta, \qquad (6c)$$

$$m = 1, \ldots, M,\ n = 1, \ldots, N,\ d = 1, \ldots, D.$$

It can be easily shown that the LME problem in (6) is bounded below (by extending the proof of Theorem 5.1 in [7]). However, we will show in Section 3.1 that the LME problem (6) is equivalent to a nonconvex indefinite quadratic optimization problem, which in general is a computationally difficult optimization problem [10]. In order to obtain a suboptimal solution in the vicinity of the global optimum, we present in Section 3.2 an approximation method for problem (6) based on convex relaxation. Thus an approximate solution of the LME problem (6) can be efficiently obtained by solving its convex relaxation counterpart.

3.1. Problem Reformulation

In the subsection, we show that the LME problem (6) is equivalent to a nonconvex quadratic optimization problem. For ease of later use, let us define the following notations

$$\begin{aligned}\tilde{\boldsymbol{\sigma}}_{m,n} &= [1/\sigma_{m,n,1}, ..., 1/\sigma_{m,n,D}]^T, \\ \tilde{\boldsymbol{\sigma}}_m &= [(\tilde{\boldsymbol{\sigma}}_{m,1})^T, ..., (\tilde{\boldsymbol{\sigma}}_{m,N})^T]^T, \tilde{\boldsymbol{\sigma}} = [(\tilde{\boldsymbol{\sigma}}_1)^T, ..., (\tilde{\boldsymbol{\sigma}}_M)^T]^T, \\ \tilde{\boldsymbol{\mu}}_{m,n} &= [\mu_{m,n,1}/\sigma_{m,n,1}, ..., \mu_{m,n,D}/\sigma_{m,n,D}]^T, \\ \tilde{\boldsymbol{\mu}}_m &= [(\tilde{\boldsymbol{\mu}}_{m,1})^T, ..., (\tilde{\boldsymbol{\mu}}_{m,N})^T]^T, \tilde{\boldsymbol{\mu}} = [(\tilde{\boldsymbol{\mu}}_1)^T, ..., (\tilde{\boldsymbol{\mu}}_M)^T]^T, \\ \mathbf{X}_{k,t} &= \mathrm{diag}\{x_{k,t,1}, ..., x_{k,t,D}\},\ (\text{a } D\times D \text{ diagonal matrix}) \\ q(m,n,d) &= (m-1)ND + (n-1)D + d, \\ \boldsymbol{S}_n^{(m)} &= \left[\mathbf{0}_{D\times q(m,n,0)},\ \mathbf{I}_D,\ \mathbf{0}_{D\times(K-q(m,n+1,0))}\right] \in \mathbb{R}^{D\times K},\end{aligned}$$

where $K = MND$, $\mathbf{0}_{D\times q(m,n,0)}$ denotes the $D \times q(m,n,0)$ zero matrix, and $\mathbf{I}_D$ is the $D \times D$ identity matrix. Then the third term on the right-hand side of (4) can be expressed as

$$\begin{aligned}\sum_{t=1}^{T}\sum_{d=1}^{D}\left(\frac{x_{k,t,d} - \mu_{k_t(m),d}}{\sigma_{k_t(m),d}}\right)^2 &= \sum_{t=1}^{T} \|\mathbf{X}_{k,t}\tilde{\boldsymbol{\sigma}}_{m,k_t(m)} - \tilde{\boldsymbol{\mu}}_{m,k_t(m)}\|^2 \\ &= \sum_{t=1}^{T} \|\mathbf{X}_{k,t}\boldsymbol{S}^{(m)}_{k_t(m)}\tilde{\boldsymbol{\sigma}} - \boldsymbol{S}^{(m)}_{k_t(m)}\tilde{\boldsymbol{\mu}}\|^2.\end{aligned}$$

[1] In the paper, we use $\{\mu_{m,n,d}\}_{d=1}^D$ and $\{\sigma^2_{m,n,d}\}_{d=1}^D$ to represent the mean and variance parameters of state n, word W_m. For the mean and variance parameters of state $k_t(m)$ of word W_m, we for notational simplicity use $\{\mu_{k_t(m),d}\}_{d=1}^D$ and $\{\sigma^2_{k_t(m),d}\}_{d=1}^D$ instead.

By defining $\tilde{\boldsymbol{y}} = [\tilde{\boldsymbol{\sigma}}^T, \tilde{\boldsymbol{\mu}}^T]^T \in \mathbb{R}^{2K}$ and $\boldsymbol{A}_k^{(m)} = (\tilde{\boldsymbol{A}}_k^{(m)})^T \tilde{\boldsymbol{A}}_k^{(m)}$,

$$\tilde{\boldsymbol{A}}_k^{(m)} = \begin{bmatrix} \mathbf{X}_{k,1}\boldsymbol{S}_{k_1(m)}^{(m)} & -\boldsymbol{S}_{k_1(m)}^{(m)} \\ & \vdots \\ \mathbf{X}_{k,T}\boldsymbol{S}_{k_T(m)}^{(m)} & -\boldsymbol{S}_{k_T(m)}^{(m)} \end{bmatrix},$$

one can recast (4) as the following quadratic form

$$\mathcal{F}(\boldsymbol{X}_k|\boldsymbol{\theta}_m) = \gamma_k(m) + \sum_{t=1}^{T}\sum_{d=1}^{D} \log(\tilde{y}_{q(m,k_t(m),d)}) - \frac{1}{2}\tilde{\boldsymbol{y}}^T \boldsymbol{A}_k^{(m)} \tilde{\boldsymbol{y}}.$$

It follows that the spherical constraints in (5) can also be reformulated in a similar manner

$$\tilde{\boldsymbol{y}}^T \boldsymbol{Q} \tilde{\boldsymbol{y}} \le r^2, \tag{7a}$$

$$\tilde{\ell}_{q(m,n,d)} \le \tilde{y}_{q(m,n,d)} \le \tilde{u}_{q(m,n,d)}, \tag{7b}$$

for $m = 1, \ldots, M,\ n = 1, \ldots, N$, and $d = 1, \ldots, D$, where

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{U}^T\boldsymbol{U} & -\boldsymbol{U} \\ -\boldsymbol{U} & \mathbf{I}_K \end{bmatrix} \in \mathbb{R}^{2K\times 2K},$$

$$\boldsymbol{U} = \mathrm{diag}\{\bar{\mu}_{1,1,1}, \bar{\mu}_{1,1,2}, \ldots, \bar{\mu}_{M,N,D}\} \in \mathbb{R}^{K\times K},$$

$\tilde{u}_{q(m,n,d)} = 1/(\bar{\sigma}_{m,n,d} - \beta) \ge 0$ and $\tilde{\ell}_{q(m,n,d)} = 1/(\bar{\sigma}_{m,n,d} + \beta)$. Therefore, we can rewrite the LME problem in (6) as

$$\min_{\tilde{\boldsymbol{y}}} \max_{\substack{m\in\Omega, m\neq i_k \\ k\in\mathcal{S}}} d_m(\boldsymbol{X}_k|\,\tilde{\boldsymbol{y}}) \tag{8a}$$

$$\text{s.t. } \tilde{\boldsymbol{y}}^T \boldsymbol{Q} \tilde{\boldsymbol{y}} \le r^2, \tag{8b}$$

$$\tilde{\ell}_{q(m,n,d)} \le \tilde{y}_{q(m,n,d)} \le \tilde{u}_{q(m,n,d)}, \tag{8c}$$

$$m = 1, \ldots, M,\ n = 1, \ldots, N,\ d = 1, \ldots, D,$$

where $d_m(\boldsymbol{X}_k|\,\tilde{\boldsymbol{y}}) = \mathcal{F}(\boldsymbol{X}_k|\boldsymbol{\theta}_m) - \mathcal{F}(\boldsymbol{X}_k|\boldsymbol{\theta}_{i_k})$ which is given by

$$d_m(\boldsymbol{X}_k|\,\tilde{\boldsymbol{y}}) = -\gamma(i_k, m) + \sum_{t=1}^{T}\sum_{d=1}^{D} \log(\tilde{y}_{q(m,k_t(m),d)}) - \sum_{t=1}^{T}\sum_{d=1}^{D} \log(\tilde{y}_{q(i_k,k_t(i_k),d)}) + \frac{1}{2}\tilde{\boldsymbol{y}}^T \boldsymbol{A}(i_k, m)\tilde{\boldsymbol{y}}, \tag{9}$$

in which $\gamma(i_k, m) = \gamma_k(i_k) - \gamma_k(m)$ and $\boldsymbol{A}(i_k, m) = \boldsymbol{A}_k^{(i_k)} - \boldsymbol{A}_k^{(m)}$. It can be observed that the minmax problem in (8) is not a convex optimization problem, because in the objective function (9) the first log term is concave and the $\boldsymbol{A}(i_k, m)$ could be indefinite. It is generally very difficult to solve a nonconvex indefinite quadratic optimization problem [10], therefore we consider an approximation method using convex relaxation in the next subsection.

3.2. Approximation by Convex Relaxation

To approximate (8) by a convex problem, we reconsider the standard deviation variables

$$\boldsymbol{\sigma}_{m,n} = [\sigma_{m,n,1}, \ldots, \sigma_{m,n,D}]^T,\ \boldsymbol{\sigma}_m = [(\boldsymbol{\sigma}_{m,1})^T, \ldots, (\boldsymbol{\sigma}_{m,N})^T]^T,$$
$$\boldsymbol{\sigma} = [(\boldsymbol{\sigma}_1)^T, \ldots, (\boldsymbol{\sigma}_M)^T]^T,$$

and define $\boldsymbol{y} = [\tilde{\boldsymbol{y}}^T, \boldsymbol{\sigma}^T, 1]^T \in \mathbb{R}^{3K+1}$ and

$$\mathbf{Y} \triangleq \begin{bmatrix} \mathbf{Y}_{11} & \mathbf{Y}_{12} \\ \mathbf{Y}_{21} & \mathbf{Y}_{22} \end{bmatrix} = \boldsymbol{y}\boldsymbol{y}^T = \left[\begin{array}{cc|cc} \tilde{\boldsymbol{\sigma}}\tilde{\boldsymbol{\sigma}}^T & \tilde{\boldsymbol{\sigma}}\tilde{\boldsymbol{\mu}}^T & \tilde{\boldsymbol{\sigma}}\boldsymbol{\sigma}^T & \tilde{\boldsymbol{\sigma}} \\ \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\sigma}}^T & \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^T & \tilde{\boldsymbol{\mu}}\boldsymbol{\sigma}^T & \tilde{\boldsymbol{\mu}} \\ \hline \boldsymbol{\sigma}\tilde{\boldsymbol{\sigma}}^T & \boldsymbol{\sigma}\tilde{\boldsymbol{\mu}}^T & \boldsymbol{\sigma}\boldsymbol{\sigma}^T & \boldsymbol{\sigma} \\ \tilde{\boldsymbol{\sigma}}^T & \tilde{\boldsymbol{\mu}}^T & \boldsymbol{\sigma}^T & 1 \end{array}\right], \tag{10}$$

where $\mathbf{Y}_{11} \in \mathbb{R}^{2K\times 2K}$, $\mathbf{Y}_{12} \in \mathbb{R}^{2K\times(K+1)}$, $\mathbf{Y}_{21} \in \mathbb{R}^{(K+1)\times 2K}$ and $\mathbf{Y}_{22} \in \mathbb{R}^{(K+1)\times(K+1)}$ are submatrices of $\mathbf{Y}$, respectively. Note that (10) is equivalent to saying that $\mathbf{Y} \succeq \mathbf{0}$ (positive semidefinite), $\mathrm{rank}(\mathbf{Y}) = 1$, and $[\mathbf{Y}_{22}]_{K+1,K+1} = 1$. By (10), we can express (9) in terms of $\mathbf{Y}$

$$d_m(\boldsymbol{X}_k|\,\mathbf{Y}) = -\gamma(i_k, m) - \sum_{t=1}^{T}\sum_{d=1}^{D}\Big(\log([\mathbf{Y}_{22}]_{q(m,k_t(m),d),K+1}) + \log([\mathbf{Y}_{12}]_{q(i_k,k_t(i_k),d),K+1})\Big) + \frac{1}{2}\mathrm{trace}(\boldsymbol{A}(i_k, m)\mathbf{Y}_{11}), \tag{11}$$

where $1/[\mathbf{Y}_{22}]_{q(m,k_t(m),d),K+1}$ and $[\mathbf{Y}_{12}]_{q(i_k,k_t(i_k),d),K+1}$ are in place of the $\tilde{y}_{q(m,k_t(m),d)}$ and $\tilde{y}_{q(i_k,k_t(i_k),d)}$ in (9), respectively. Hence $d_m(\boldsymbol{X}_k|\,\mathbf{Y})$ is convex on $\mathbf{Y}$. The use of $\boldsymbol{\sigma}$ leads to the nonconvex constraints $\sigma_i\tilde{\sigma}_i = 1,\ i = 1, \ldots, K$, which, by (10), are equivalent to

$$[\mathbf{Y}_{12}]_{i,i} = 1,\ i = 1, \ldots, K. \tag{12}$$

Therefore, we can conclude from (10), (11), and (12) that the LME problem in (8) is equivalent to the following problem

$$\min_{\mathbf{Y}} \max_{\substack{m\in\Omega, m\neq i_k \\ k\in\mathcal{S}}} d_m(\boldsymbol{X}_k|\,\mathbf{Y}) \tag{13a}$$

$$\text{s.t. } \mathrm{trace}(\boldsymbol{Q}\mathbf{Y}_{11}) \le r^2, \tag{13b}$$

$$\tilde{\ell}^2_{q(m,n,d)} \le [\mathbf{Y}_{11}]_{q(m,n,d),q(m,n,d)} \le \tilde{u}^2_{q(m,n,d)}, \tag{13c}$$

$$\ell^2_{q(m,n,d)} \le [\mathbf{Y}_{22}]_{q(m,n,d),q(m,n,d)} \le u^2_{q(m,n,d)}, \tag{13d}$$

$$\tilde{\ell}_{q(m,n,d)} \le [\mathbf{Y}_{12}]_{q(m,n,d),K+1} \le \tilde{u}_{q(m,n,d)}, \tag{13e}$$

$$\ell_{q(m,n,d)} \le [\mathbf{Y}_{22}]_{q(m,n,d),K+1} \le u_{q(m,n,d)}, \tag{13f}$$

$$m = 1, \ldots, M,\ n = 1, \ldots, N,\ d = 1, \ldots, D,$$

$$[\mathbf{Y}_{12}]_{i,i} = 1,\ i = 1, \ldots, K, \tag{13g}$$

$$[\mathbf{Y}_{22}]_{K+1,K+1} = 1,\ \mathbf{Y} \succeq \mathbf{0}, \tag{13h}$$

$$\mathrm{rank}(\mathbf{Y}) = 1,$$

where (13e) and (13f) are due to (5b) and (8c), $u_{q(m,n,d)} = \bar{\sigma}_{m,n,d} + \beta$, and $\ell_{q(m,n,d)} = \bar{\sigma}_{m,n,d} - \beta \ge 0$. Note that we also have imposed the constraint (5b) to the diagonal entries of $\mathbf{Y}_{11}$ and $\mathbf{Y}_{22}$, leading to (13c) and (13d), respectively. Finally, by discarding the rank one constraint $\mathrm{rank}(\mathbf{Y}) = 1$, we obtain a convex relaxation counterpart of problem (6).

Once the optimum solution $\mathbf{Y}^\star$ of the relaxation problem (13) (without the constraint $\mathrm{rank}(\mathbf{Y}) = 1$) is obtained, we need to approximate the large-margin CDHMM solutions of problem (6) from $\mathbf{Y}^\star$. An approximate solution can be obtained as follows

$$\hat{\sigma}_{m,n,d} = \frac{1}{\sqrt{[\mathbf{Y}^\star_{11}]_{q(m,n,d),q(m,n,d)}}}, \tag{14a}$$

$$\hat{\mu}_{m,n,d} = \mathrm{sign}([\mathbf{Y}^\star_{11}]_{q(m,n,d),K+q(m,n,d)}) \times \sqrt{[\mathbf{Y}^\star_{11}]_{K+q(m,n,d),K+q(m,n,d)}}\,\hat{\sigma}_{m,n,d}, \tag{14b}$$

by assuming that the obtained $\mathbf{Y}^\star$ is of rank one with the same structure as in (10).

4. SIMULATION RESULTS AND DISCUSSIONS

In the section, we present some simulation results to demonstrate the efficacy of the presented LMMVE. In the simulation, we represent each speech "unit" (such as a word) by a 3-state ($N = 3$), left-to-right CDHMM with a 4×1 multivariate Gaussian vector ($D = 4$) for each state. That is, the HMM is specified by the

parameter set $\boldsymbol{\theta} = \{\{a_{i,n}\}_{i,n=1}^{3}, \{\pi_n\}_{n=1}^{3}, \{\boldsymbol{\mu}_n\}_{n=1}^{3}, \{\boldsymbol{\sigma}_n\}_{n=1}^{3}\}$, where $\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n \in \mathbb{R}^D$, $\pi_1 = 1$, $\pi_n = 0$ for $n = 2, 3$, and $a_{i,n} = 0$ for $i < n$. A set of 10 HMMs ($M = 10$), each of which represents a speech unit, was simulated, and each HMM differs from others by different set of parameters. The parameters of each HMM were randomly generated. Then with these generated parameters, a total number of 1000 training signals (100 training signals labeled by each HMM) were randomly generated. The length T of each (training and testing) signal was randomly selected from 30 to 50, simulating speech signals with variable lengthes for a fixed speech unit. To simulate the model mismatch problem, we generate the testing signals $\tilde{\boldsymbol{X}} = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_T\}$ as follows. Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ be a generated CDHMM signal. In order to simulate some well-known properties of speech signals [12, 13], we introduced correlation into the signal itself by $\tilde{x}_{t,d} = \sum_{\ell=0}^{\infty} \alpha^{\ell} u[\ell] x_{t-\ell,d}$ for $d = 1, \ldots, D$ and $t = 1, \ldots, T$, where $u[t]$ is the unit step sequence, and parameter $0 < \alpha \leq 1$ decides the degree of the model mismatch from CDHMM. There were over 5000 testing signals (500 signals labeled by each HMM) generated in the simulation. The LME algorithm proposed in [8] was considered in the simulation. The Baum-Welch MLE [1] was first applied, and the obtained ML estimates were used as the initial model of the LME algorithm as well as the search center model $\{\bar{\mu}_{m,n,d}, \bar{\sigma}^2_{m,n,d}\}$ of the LMMVE (see (5)). The 150 training signals which were relatively closest to the decision boundary of ML estimates were selected to form the support token set $\mathcal{S}$ (in (2)). The relaxation problem (13) was solved by a specially developed first-order convex optimization algorithm [14]. We compared the proposed LMMVE with the MLE [1] and the SDP based LMME [7].

In our simulation, all three estimators achieved 100% recognition rate on the training signal set. Table 1 lists the minimum and average margin values of training signals (see (1)) associated with the MLE estimates, the LMME estimates and the LMMVE estimates, respectively. One can see from the table that both the LMME and the LMMVE have larger margin values than the MLE, but the LMMVE has the highest ones. Figure 1 shows the testing recognition rates of estimators under model mismatch for $\alpha = 0.1, 0.2, 0.3, 0.4$, and 0.5. It can be seen from the figure that, with increased degree of model mismatch the LMMVE exhibits much better robustness compared to the MLE and the LMME, demonstrating the advantage of the joint mean-and-variance estimation method.

Table 1. Margin values of estimators.

	Minimum value	Average value
MLE	97.2	367.4
LMME ($r = 3$)	202.4	422.5
LMMVE ($r = 3, \beta = 0.2$)	245.8	558.6
LMME ($r = 5$)	199.3	460.8
LMMVE ($r = 5, \beta = 0.2$)	239.5	594.5

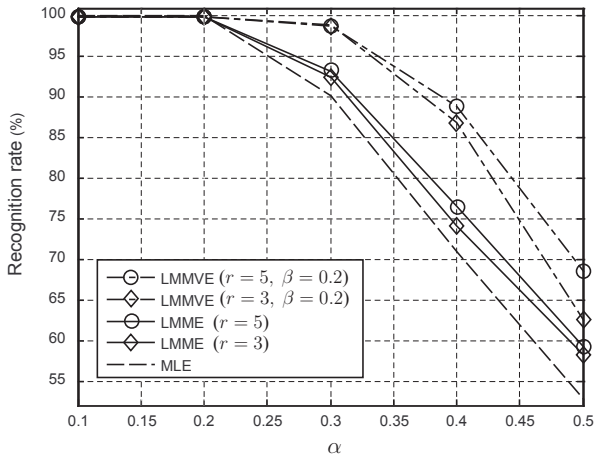

Figure 1. Recognition rates (%) of estimators under model mismatch.

5. CONCLUSIONS

In the paper, we have presented a convex relaxation based parameter estimation technique for large-margin CDHMM. The proposed technique is the first attempt to jointly estimate the means and variances in normalized Gaussian CDHMMs with the large-margin optimization criterion. We have shown that the associated estimation problem is a computationally difficult optimization problem, but can be efficiently approximated by a convex relaxation problem. Our simulation results have demonstrated the benefit of the joint estimation technique and the effectiveness of the presented approximation method.

6. REFERENCES

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *IEEE Proceedings*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

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

[3] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Cambridge University, Cambridge, UK, 2004.

[4] V. Vapnik, *Statistical Learning Theory*. New York: John-Wiley & Sons, 1998.

[5] A. J. Smola, P. L. Bartlett, B. Scholkopf, and D. Schuurmans, *Advances in Large Margin Chassifiers*. Cambridge, Massachusetts: The MIT Press, 2000.

[6] F. Sha and L. K. Saul, "Large margin hidden markov models for automatic speech recognition," in *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, MA, 2007.

[7] X. Li, "Large margin hidden Markov models for speech recognition," Master's thesis, Graduate Program in Computer Science, York University, Toronto, Ontario, Canada, Sept. 2005.

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

[9] H. Jiang and X. Li, "Incorporating tracking errors for large margin HMMs under semidefinite programming framework," in *Proc. ICASSP*, Honolulu, Hawaii, April 15-20, 2007, pp. IV629–IV632.

[10] S. He, Z.-Q. Luo, J. Nie, and S. Zhang, "Semidefinite relaxation bounds for indefinite homogeneous quadratic optimization," submitted to *SIAM J. Optimization*.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.

[12] H. Sheikhzadeh and L. Deng, "Waveform-based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization," *IEEE Trans. Speech and Audio Process.*, vol. 2, no. 1, pp. 80–91, January 1994.

[13] L. Deng, "A stochastic model of speech incorporating hierarchical nonstationarity," *IEEE Trans. Speech and Audio Process.*, vol. 1, no. 4, pp. 471–475, Oct. 1993.

[14] T.-H. Chang, Z.-Q. Luo, L. Deng, and C.-Y. Chi, "Parameter estimation of large-margin continuous-density HMM by convex relaxation," in preparation.