Large-Margin Discriminative Training of Hidden Markov Models for Speech Recognition

Dong Yu and Li Deng
Microsoft Research, One Microsoft Way, Redmond WA, 98052-6399, USA
{dongyu, deng}@microsoft.com

Abstract

Discriminative training has been a leading factor for improving automatic speech recognition (ASR) performance over the last decade. The traditional discriminative training, however, has been aimed to minimize empirical error rates on training sets, which may not be well generalized to test sets. Many attempts have been made recently to incorporate the principle of large margin (PLM) into the training of hidden Markov models (HMMs) in ASR to improve the generalization abilities. Significant error rate reduction on the test sets has been observed on both small vocabulary and large vocabulary continuous ASR tasks using large-margin discriminative training (LMDT) techniques. In this paper, we introduce the PLM, define the concept of margin in the HMMs, and survey a number of popular LMDT algorithms proposed and developed recently. Specifically, we review and compare the large-margin minimum classification error (LM-MCE) estimation, soft-margin estimation (SME), large margin estimation (LME), large relative margin estimation (LRME), and large margin training (LMT) with a focus on the insights, the training criteria, the optimization techniques used, and the strengths and weaknesses of these different approaches. We suggest future research directions in our conclusion of this paper.

1. Introduction

Traditionally, the parameters of hidden Markov models (HMMs) in automatic speech recognition (ASR) systems are learnt from speech corpora using the maximum likelihood estimation (MLE) [2], which maximizes the joint likelihood over utterances and their transcriptions in the training data. MLE is appealing for two reasons. First, in the asymptotic limit of infinite training data, MLE provides a minimum-variance and consistent estimate of the true parameters for the distributions when the model is accurate. Second, MLE can be efficiently carried out using the expectation-maximization (EM) and forward backward training algorithms [2]. Note, however, that MLE does not directly minimize word or phoneme recognition error rates and often does not provide the best decision boundaries in terms of minimizing the error rates, since HMMs are only crude approximate models of speech and the *a priori* probabilities of the word sequence estimated using the stochastic language models are grossly inexact also.

Knowing the weaknesses of MLE, researchers have proposed using discriminative training criteria that are more closely related to the recognition error rates than the MLE. The most notable discriminative training criteria include maximum mutual information (MMI) [5], minimum classification error (MCE) [4], minimum word error (MWE), and minimum phone error (MPE) [18]. Adoption of these discriminative training criteria for estimating HMM parameters has been a leading factor in decreasing the ASR error rates in large vocabulary ASR tasks over the last decade. Traditional discriminative training such as MMI, MCE, and MPE aims to find classification boundaries that minimize empirical error rates on training sets, which may not be well generalized to test sets.

Recently, many attempts have been made to incorporate the principle of large margin (PLM) into the training of HMMs in ASR to improve the generalization abilities. Significant error rate reduction over the traditional discriminative training on the test sets has been observed on both small vocabulary (e.g., TIDIGITS and TIMIT) and large vocabulary continuous ASR tasks (e.g., Wall Street Journal and large vocabulary telephony applications) using large margin discriminative training (LMDT) techniques.

The PLM has been intensively investigated by researchers of the statistical learning community to balance the empirical error rate on the training set and the generalization ability on the test set. Intuitively, a classifier with larger margin can better tolerate the mismatches between the training and test sets. Vapnik

0-7695-2997-6/07 $25.00 © 2007 IEEE
DOI 10.1109/ICSC.2007.11

[17] has shown that the test-set error rate is bounded by the sum of the empirical error rate on the training set and a generalization score associated with the margin. Minimizing this bound can better achieve the goal of minimizing the test set error rate than minimizing the empirical training set error alone.

The PLM has been successfully used in designing the state-of-the-art multi-way classifiers (most notably support vector machines (SVMs) [17]) for many years. Its application to estimating the HMM parameters for ASR, however, is more recent due to the challenges introduced when extending the PLM from multi-way classification to sequential classification. The two mostly cited challenges are the formation of the training criteria and the derivation of efficient optimization algorithms.

In this paper we define the concept of margin in the HMMs, and survey the LMDT algorithms for estimating HMM parameters. Specifically, we review and compare the large-margin minimum classification error (LM-MCE) [19] [20] estimation, soft-margin estimation (SME) [6] [7], large margin estimation (LME) [3] [8] [9] [10], large relative margin estimation (LRME) [11][12], and large margin training (LMT) [15] [16] with the focus on the insights, the training criteria, the optimization techniques used, and strengths and weaknesses of these various approaches.

The rest of the paper is organized as follows. In section 2, we introduce HMM - the standard ASR acoustic model (AM) and define the margin in HMM. In section 3, we describe the LM-MCE and SME techniques, both of which are direct extensions to the traditional discriminative training algorithms by including margin in the training criterion and have been successfully applied to large vocabulary continuous ASR tasks. In section 4, we review the LME and LRME whose optimization can be efficiently carried out based on the SDP [1] formulation of the problem. Contrast to the LM-MCE and SME techniques, LME and LRME techniques were originally proposed to maximize the margin and later extended to include the error rates in the training criterion. LME was the first attempt to incorporate the PLM into the discriminative training of HMM parameters. In section 5, we describe LMT whose key idea was borrowed from the conjugate form of the SVM. In section 6, we summarize these LMDT algorithms and suggest future research directions.

2. Margin in hidden Markov models

Speech recognition is a sequential classification problem which is significantly more difficult than the non-sequential multi-way classification problem. In this section, we introduce HMM – the standard AM in ASR and define the margin in HMM.

2.1. HMM

In the state-of-the-art ASR systems, speech units (e.g. phonemes) are typically modeled by an S-state continuous density HMM (CDHMM) with parameter set $\lambda = (\tau, \mathrm{A}, \theta, \mathcal{L})$, where the set of hidden states is denoted as $\{0, 1, 2, \ldots, \mathrm{S}, \mathrm{S}+1\}$ with state 0 and S+1 being the augmented entry and exit states (both emit null observation) respectively, $\mathcal{L}$ is the language model (LM), $\tau(\mathrm{s})$ is the probability that the system is initially in the state s, $\mathrm{A} = \{\mathrm{a_{ij}}\}_{\mathrm{i,j=1}}^{\mathrm{S}}$ is the transition probability matrix where $\mathrm{a_{ij}}$ is the probability of transferring from state $s_t = i$ at time t to state $s_{t+1} = j$ at time $\mathrm{t}+1$, and θ is the state observation probability parameters composed of multivariate Gaussian mixture distributions with parameters $\theta_\mathrm{i} = \{\omega_\mathrm{ik}, \mu_\mathrm{ik}, \Sigma_\mathrm{ik}\}_{\mathrm{k=1}}^{\mathrm{K}}$ for each state i, with K being the number of Gaussian mixtures in each state. The state observation probability density function (pdf) is usually assumed to have diagonal covariance matrices and can be estimated as

$$\begin{aligned} \mathrm{b_i}(x_t) &= p(x_t | s_t = i\,; \theta) \\ &= \sum_{k=1}^{K} \omega_{ik} \mathcal{N}(x_t; \mu_\mathrm{ik}, \Sigma_\mathrm{ik}) \\ &= \sum_{k=1}^{K} \omega_{ik} \prod_{d=1}^{D} \sqrt{\frac{1}{2\pi\sigma_{ikd}^2}}\, e^{-\frac{(x_{td}-\mu_\mathrm{ikd})^2}{2\sigma_{ikd}^2}}, \end{aligned} \tag{1}$$

where mixture weights ω_{ik}'s satisfy the constraint

$$\sum_{k=1}^{K} \omega_{ik} = 1, \tag{2}$$

and $\Sigma_\mathrm{ik} = diag(\sigma_{ik1}^2, \sigma_{ik2}^2, \ldots, \sigma_{ikD}^2)$ is the diagonal covariance matrix of k-th Gaussian in state i.

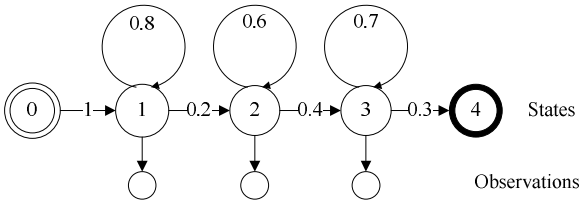

Figure 1. An example of HMM

An example of 3-state HMM (with a left-to-right topology) is illustrated in Figure 1, where states 0 and 4 are the augmented entry and exit states. In the continuous speech recognition systems, phoneme HMMs are connected (by linking the exit state of the previous phone to the entry state of the next phone) into word HMMs with phoneme transition probabilities

determined upon the word pronunciation model, and word HMMs are connected into sentence HMMs with word transition probabilities estimated using the LM.

Given a speech utterance $X = \{x_t\}_{t=1}^{T}$ the recognition result is determined by the discriminant function of the following form:

$$\begin{aligned} \mathcal{F}(X, W|\lambda) &= \log P(W|X;\lambda) \\ &\propto \log[P(X|W;\lambda)P(W|\lambda)] \\ &= \log P(X|W;\lambda) + \log P(W|\lambda), \end{aligned} \tag{3}$$

where W is a word sequence, $P(W|\lambda)$ is the LM score of W and

$$\begin{aligned} p(X|W;\lambda) &= \sum_{s=1}^{S} p(X, s|\lambda) \\ &= \sum_{s} \tau(s_1) \prod_{t=2}^{T} a_{s_{t-1}s_t} \prod_{t=1}^{T} b_{s_t}(x_t) \end{aligned} \tag{4}$$

is the AM score where $s = \{s_t\}_{t=1}^{T}$ is the state sequence. The AM score (4) may be approximated using the score along the best state sequence s^* as

$$p(X|W;\lambda) \cong \max_{s} \tau(s_1) \prod_{t=2}^{T} a_{s_{t-1}s_t} \prod_{t=1}^{T} b_{s_t}(x_t). \tag{5}$$

The recognized word sequence $\widehat{W}$ is the one that maximizes the discriminant function, i.e.,

$$\widehat{W} = \max_{W} \mathcal{F}(X, W|\lambda). \tag{6}$$

2.2. Definition of margin in HMM

Margin in HMM is defined differently for the training set than that for the test set. For each utterance X in the test set, the margin $m(X|\lambda)$ is defined as the distance between the discriminant score of the recognized word sequence and the discriminant score of the closest competing word sequence, or

$$\begin{aligned} m(X|\lambda) &= \min_{W \neq \widehat{W}} \left[\mathcal{F}\left(X, \widehat{W}|\lambda\right) - \mathcal{F}(X, W|\lambda)\right] \\ &= \mathcal{F}\left(X, \widehat{W}|\lambda\right) - \max_{W \neq \widehat{W}} \mathcal{F}(X, W|\lambda). \end{aligned} \tag{7}$$

The margin of an utterance in the test set is always greater than or equal to 0 and is a key feature in deriving the confidence score of the recognition result.

For each utterance X in the training set, on the other hand, the margin is defined as the distance between the discriminant score of the label $\overline{W}$ to the highest discriminant score of all the competing word sequences, i.e.,

$$\begin{aligned} m(X|\lambda) &= \min_{W \neq \overline{W}} [\mathcal{F}(X, \overline{W}|\lambda) - \mathcal{F}(X, W|\lambda)] \\ &= \mathcal{F}(X, \overline{W}|\lambda) - \max_{W \neq \overline{W}} \mathcal{F}(X, W|\lambda). \end{aligned} \tag{8}$$

Note that the margin of an utterance in the training set may be less than zero (as shown in Figure 2) since the discriminant score of the label may not be the highest among all the discriminant scores.

The margin of the entire training set $\Omega = \{X_i\}_{i=1}^{N}$ is defined as the minimal margin of all the correctly classified utterances, i.e.,

$$m(\Omega|\lambda) = \min_{X_i : m(X_i|\lambda) \geq 0} m(X_i|\lambda). \tag{9}$$

The margin of the training set is always greater than or equal to zero since misclassified training utterances are not used in calculating the margin. Note that the classifier with a larger training-set margin sometimes may have a larger training-set error rate than the classifier with a smaller training set margin. For example, both the margin and the training-set error rate defined by the solid line in Figure 2 are larger than that defined by the dashed line.

According to the above definitions, margins are functions of the model parameters λ. By adjusting the model parameters, we can alter the margins. In this paper, we assume the language model is fixed and we are only interested in estimating HMM parameters using LMDT techniques.

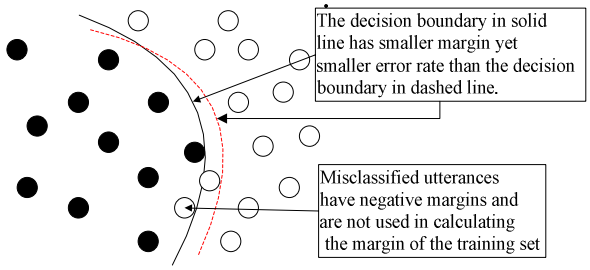

Figure 2. Margins defined in the training set

3. LM-MCE and SME

Traditional discriminative training algorithms have been aimed to minimize some measure of empirical error rates in the training set. One obvious way of applying the PLM in the training process is to extend these algorithms by optimizing a combined score of the training set error rate and the margin. In this section we introduce two approaches of this kind: LM-MCE [19] [20] and SME [6] [7], both of which have been successfully applied to large-vocabulary continuous ASR.

3.1. LM-MCE

LM-MCE is an extension to and reinterpretation of the traditional MCE [4], which aims to minimize the empirical sentence error rate (SER)

$$\mathcal{R}^{MCE}(\lambda) = \frac{1}{N} \sum_{i=1}^{N} r^{MCE}(X_i)$$

$$= \frac{1}{N}\sum_{i=1}^{N} g\big(-m(X_i|\lambda)\big) \tag{10}$$

in the training set, where

$$g(x) = \delta(x \geq 0) = \begin{cases} 1 & if\ x \geq 0 \\ 0 & else \end{cases} \tag{11}$$

is the step function. Since $g(x)$ is not differentiable, it is usually replaced by the sigmoid function

$$f(x;\alpha,\beta) = \frac{1}{1+e^{-\alpha(x+\beta)}} \tag{12}$$

where β is usually set to zero and

$$\lim_{\alpha\to\infty} f(x;\alpha,\beta) = g(x). \tag{13}$$

As a result, the MCE training selects the model parameters such that

$$\hat{\lambda} = \arg\min_{\lambda} \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1+e^{\alpha m(X_i|\lambda)}}. \tag{14}$$

To incorporate PLM in the training criterion, Yu et al. [19] [20] defined a new risk function

$$r^{LM-MCE}(X_i|\lambda) = g\big(-(m(X_i|\lambda)-\rho)\big) \tag{15}$$

for each utterance, where $\rho > 0$ is the minimum margin required (MMR) for an utterance in the training set to be considered correctly classified. LM-MCE is aimed to minimize the smoothed risk

$$\mathcal{R}^{LM-MCE}(\lambda) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1+e^{\alpha(m(X_i|\lambda)-\rho)}}. \tag{16}$$

Note that this is a simple extension to the MCE training by choosing $\beta = \rho$ in the sigmoid function.

The insights of the extension can be gained if we separate the above risk into two terms

$$\begin{aligned} \mathcal{R}^{LM-MCE}(\lambda) &= \frac{1}{N}\sum_{i=1}^{N} g\big(-(m(X_i|\lambda)-\rho)\big) \\ &= \frac{1}{N}\sum_{i=1}^{N} g\big(-m(X_i|\lambda)\big) \\ &\quad + \frac{1}{N}\sum_{i=1}^{N} \delta(0 < m(X_i|\lambda) \leq \rho). \end{aligned} \tag{17}$$

The first term is the sentence error rate and the second term is a function of MMR ρ.

The MCE criterion smoothed by the sigmoid function has been shown by McDermott and Katagiri [14] to be equivalent to minimizing the estimated empirical test set error rate. In other words, the conventional MCE (with $\beta = 0$) has some built-in generalization ability. Yu et al. [20] extended this result and showed that by setting a positive β in the sigmoid function, LM-MCE training optimizes a combined risk on the test set. In addition, strong empirical ASR results were reported in [19][20] after parameter β was carefully selected from iteration to iteration in the MCE-style training.

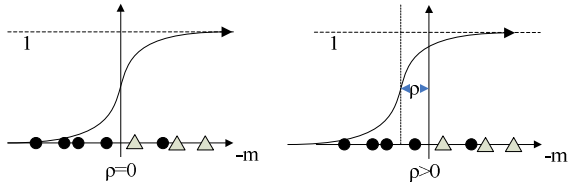

Figure 3. Illustration of the LM-MCE training

In extending the MCE training criterion, LM-MCE inherits two advantages from MCE. First, LM-MCE training is not sensitive to the outliers since the sigmoid function saturates on extremely large values of the input. Second, many efficient optimization techniques developed for the MCE training (e.g., generalized probabilistic descent (GPD) and extended Baum Welch (EBW)) can be directly used in the LM-MCE training. The LM-MCE training, however, also bears the same weaknesses as the MCE training. For example, both methods optimize the SER on the training set, which is not effective when training ASR systems with long sentences.

Choosing a suitable MMR ρ in LM-MCE is not trivial though. In practice, it is selected using some sort of cross verification. Furthermore, using a large, fixed margin may cause the training algorithm to treat additional utterances as outliers and thus hurt the training performance. This is illustrated in Figure 3, where utterances represented with circles and triangles belong to two different classes. In the left sub-figure, MMR is set to 0 while in the right sub-figures MMR is set to a positive value. Note that circles and triangles belong to two different classes. In the left sub-figure MMR is set to zero while in the right sub-figure MMR is set to a positive value. Note that the utterance represented by the right-most circle is not treated as an outlier in the left sub-figure but is likely to be treated as an outlier in the right sub-figure since the sigmoid function saturates at that point. To solve this deficiency, Yu et al. [19] proposed to systematically increase the MMR by setting MMR to zero or even negative initially and increasing it gradually over iterations. The training process stops when the minimum word error rate (WER) on the cross validation set is achieved. The LM-MCE criterion is not convex and the training is subject to the local minimum problem.

WER of 0.19% and SER (sentence error rate) of 0.55% have been obtained on the TIDIGITS corpus using the LM-MCE training [19]. This is a 17.39%

relative WER and 19.52% relative SER reduction over the state-of-the-art MCE training. Positive results have also been reported on the large vocabulary telephony applications with more than 2000 hours of training data [20]. In that task, LM-MCE achieved 1.644% WER reduction, which translates to 16.57% relative WER reduction over the MLE and 5.64% relative WER reduction over the standard MCE training.

3.2. SME

Similar to LM-MCE, SME [6] [7] is an extension to the traditional discriminative training approaches by optimizing a combined score of the margin and some measure of empirical risk. SME is aimed to minimize

$$\mathcal{R}^{SME}(\lambda) = \frac{\epsilon}{\rho} + \frac{1}{N}\sum_{i=1}^{N} h\big(-(m'(X_i|\lambda) - \rho)\big), \quad (18)$$

where

$$h(x) = \begin{cases} x & if\ x \geq 0 \\ 0 & else \end{cases} \quad (19)$$

is the hinge function, and ϵ is a coefficient to balance soft margin maximization and error rate minimization. A smaller ϵ corresponds to a higher penalty for the empirical risk. Note that the MMR ρ in (18) above is not enforced (i.e., the right term can be greater than zero) during the optimization process. Hence, ρ is called soft-margin by the authors of the SME papers [6] [7].

Although the SME training criterion defined by (18) is very similar to the LM-MCE one, there are three differences between them. First, there is an additional term of ϵ/ρ in the SME criterion. This term decreases as the MMR ρ increases and is in favor of larger MMRs. Second, the hinge function instead of the step function is used as the measure of the empirical risk. Since the hinge function is unbounded in the positive half space, SME may be sensitive to the outliers if $m'(X_i|\lambda)$ is not well defined. On the positive side, the usage of the hinge function allows for the adoption of fine risks (instead of 0-1 risk at the sentence level as in the LM-MCE) in the SME. Third, the margin $m'(X_i|\lambda)$ in the SME can be (and usually is) different from the margin $m(X_i|\lambda)$ defined in (8). For example, in [7] it was proposed to use frame-level (i.e., only counting the frames with different phoneme IDs), phoneme-level (corresponding to MPE), word-level (corresponding to MWE), and sentence-level (corresponding to MCE) discriminant scores as $m'(X_i|\lambda)$. A list of different $m'(X_i|\lambda)$ can be found in the Table 2 of [7].

Similar to LM-MCE, SME can optimize the training criterion using the GPD training algorithm. Note that the value of the SME criterion can be changed by adjusting ρ, ϵ, or the HMM parameters. No principled way of choosing ρ and ϵ has been found in the literature. The training as reported in [7] is thus carried out in epochs by selecting ρ and ϵ heuristically, fixing them, and then adjusting the HMM parameters to minimize $\mathcal{R}^{SME}(\lambda)$. ρ and ϵ are then changed and a new epoch starts. From this point of view, SME is very similar to LM-MCE since the term ϵ/ρ is a fixed value when ρ and ϵ are fixed and both ρ and ϵ are not functions of the HMM parameters. Within each epoch only utterances that have smaller separation scores than MMR ρ contribute to the optimization process. Figure 4 depicts a state where the utterances with margins $m'(X_i|\lambda) \leq \rho$ are marked with arrowed lines whose length indicates the contribution to the object function. The training algorithm will adjust the HMM parameters to pull utterances marked as solid circles to the left of the left dotted line and the utterances marked as hollow circles to the right of the right dotted line. Note that after adjusting the HMM parameters, some utterances satisfying $m'(X_i|\lambda) > \rho$ originally may now have $m'(X_i|\lambda) \leq \rho$. This fact indicates that after one or several iterations (within the same epoch) re-decoding and alignment is necessary to train SME models. Since utterances with $m'(X_i|\lambda) > \rho$ does not have the effect in adjusting the model parameters, the SME training might be slow. Like LM-MCE, the SME criterion is not convex and is hence subject to the problem of local minimum. Finally, note that in Figure 4 we include an outlier utterance (e.g., caused by mislabeling), which isshown as the right most solid circle. This outlier utterance may dominate the SME training and lead to a poor model. In contrast, the LM-MCE training does not suffer from outlier utterances due to the use of the sigmoid function.

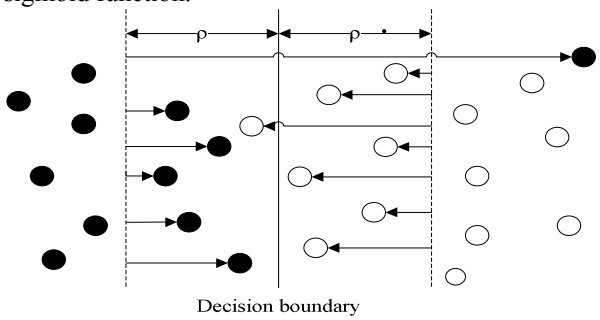

Figure 4. Utterances used in SME training

SME has been successfully applied to TIDIGITS, where a 0.67% SER [6] has been achieved. It is also successfully applied to the more challenging 5k-WSJ0 task where a 5.60% WER [7] was obtained using the word level SME and a trigram LM. This translates to an 8.6% relative WER reduction over MLE.

4. LME and LRME

Unlike LM-MCE and SME, LME [3] [8] [9] [10] and LRME [11][12] were originally designed to maximize the margins of the training set and were extended later to include the empirical error rates.

4.1. Margin only criteria

LME [8] [9] [10], in its original form, is aimed to maximize the margin of the training set

$$\mathcal{R}^{LME}(\lambda) = m(\Omega|\lambda) = \min_{X_i: m(X_i|\lambda)\geq 0} m(X_i|\lambda). \quad (20)$$

Note that in this criterion the misclassified utterances are not counted as discussed in section 2.2. Note also that this criterion is unbounded and is thus not a suitable training criterion by its own. As an example to illustrate the problem, $\mathcal{R}^{LME}(\lambda)$ can be arbitrarily increased by decreasing all the covariance matrices of the Gaussian mixtures.

To remedy this deficiency, Liu et al. proposed LRME [11] [12] which aims to maximize the relative margin

$$\mathcal{R}^{LRME}(\lambda) = \min_{X_i: m(X_i|\lambda)\geq 0} \frac{m(X_i|\lambda)}{\mathcal{F}(X_i, \overline{W}_i|\lambda)}. \quad (21)$$

Note that LRME criterion is non-convex and the training can be carried out using GPD.

An alternative approach is to formulate LME as a constrained optimization problem so that the model parameters cannot be arbitrarily changed (and so the criterion is bounded). Li et al. [8] [10] have proposed to constrain the parameters based on Kullback–Leibler (KL) divergence

$$\sum_{j=1}^{M} D\left(\lambda_j \parallel \lambda_j^{(0)}\right) \leq r^2 \quad (22)$$

between new model parameters and their initial values, where r is the bound.

Note that all these adjusted criteria have two shortcomings in common. First, they maximize only the margin and do not take into account the empirical error rate in the training set. Hence,they will work only if the training set error rate is extremely low. Second, since neither LME nor LRME training criteria include misclassified utterances. an extremely bad model with high error rates may lead to a high objective function score and be selected by LME and LRME. In other words, the quality of the initial model has significant effects on the training result. Figure 5 depicts an example of this drawback. In this example, Classifiers1 is very bad with apparently high error rates, yet it has higher margin (as measured by either (20) or (21)) than Classifier 2, which has no training-set errors. If the initial model is sufficiently good, LME and LRME will lead to an even better model since in the process of increasing the margin on the correctly classified utterances, originally misclassified utterances may become correctly classified and will contribute to the objective function. LME and LRME are not sensitive to outliers since no outliers are counted in the criteria.

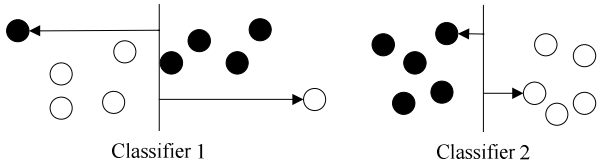

Figure 5. Examples showing deficiency of LME/LRME criteria

4.2. Soft-LME

As discussed in Section 4.1, maximizing the margin alone is not desirable under some conditions. For this reason Jiang and Li [3] recently proposed soft-LME which maximizes the combined score

$$\mathcal{R}^{soft-LME}(\lambda) = m(\Omega|\lambda) - \epsilon \cdot \frac{1}{E}\sum_{k=1}^{E} \xi(X_k|\lambda), \quad (23)$$

where $\epsilon > 0$ is a pre-set positive constant to balance the contribution from the margin of the training set and the average error in the misclassified set

$$\mathcal{E} = \{X_k | X_k \in \Omega \text{ and } m(X_k|\lambda) < 0\}_{k=1}^{E} \quad (24)$$

subject to constraint (22). This constrained optimization problem can be reformulated as

$$\hat{\lambda} = \underset{\lambda,\rho}{\operatorname{argmin}} \left[-\rho + \epsilon \cdot \frac{1}{E}\sum_{k=1}^{E} \xi(X_k|\lambda)\right] \quad (25)$$

subject to

$$\mathcal{F}(X_i, W|\lambda) - \mathcal{F}\left(X_i, \widehat{W}_i|\lambda\right) \leq -\rho,$$
$$\sum_{j=1}^{M}\sum_{d=1}^{D} \frac{\left(\mu_{jd} - \mu_{jd}^{(0)}\right)^2}{2\sigma_{jd}^2} \leq r^2, \quad (26)$$
$$\rho \geq 0.$$

Similar to other large-margin training methods, the soft-LME training can be carried out using the GPD algorithm. In fact, GPD was the optimization method used when LME/LRME were first introduced [8] [9] [11] [12]. Recently, a better optimization method has been found by Jiang and Li [10] [3]. The central idea of their approach is to formulate soft-LME as an SDP problem which is convex and can be solved using existing efficient optimization algorithms and solvers [1]. We refer the interested readers to [3] for detailed steps involved in formulating soft-LME into an SDP

problem. Here we want to point out some of the limitations in the current soft-LME/SDP algorithm. First, due to constraint (22), the optimization of both LME and soft-LME needs to be carried out in epochs, with previous epoch's result as the initial model of the new epoch. Within each epoch, the model difference is constrained by (22) and soft-LME/SDP is convex. However, the optimization is not convex across epochs and is subject to local minimum. Second, to formulate soft-LME into an SDP problem, the error term $\xi(X_k|\lambda)$ needs to be either

$$\xi(X_k|\lambda) = \frac{1}{N}\sum_{i=1}^{N}[\mathcal{F}(X,\bar{W}_k|\lambda) - \mathcal{F}(X,W_i|\lambda)] \quad (27)$$

or

$$\xi(X_k|\lambda) = -m(X_k|\lambda). \quad (28)$$

In addition, during the reformulation process, the constraint needs to be relaxed and hence it may hurt the training performance. Third, at the time of this writing, only the adaptation of the Gaussians means in HMMs can be formulated as an SDP problem. Hence, covariance matrices in HMMs are not adjusted. The effects of ignoring the LME training of covariance matrices are unknown. Fourth, the number of constraints in (26) is huge when the training set size is large. This has restricted the application of the soft-LME/SDP algorithm to large vocabulary ASR tasks since existing SDP optimization algorithms execute in batch mode and usually cannot handle large optimization problems normally seen in ASR.

Even with these limitations, the soft-LME/SDP has been shown to be very effective compared to the GPD algorithm on small tasks such as TIDIGITS where a 0.20% WER [10] [3] has been achieved using the soft-LME/SDP. Jiang and Li [3] noticed that when the training set error rate is large, soft-LME/SDP performs better than the LME/SDP. If the training-set error rate is extremely low, the soft-LME/SDP algorithm does not provide performance gain compared with LME/SDP. Under both conditions, however, soft-LME/SDP converges to the best value of the objective function much faster than LME/SDP.

5. LMT

Contrast to LME/LRME, which was motivated by the conjugate form of the SVM [17], LMT [15] [16] was motivated by the primal form of the SVM. LMT is also formulated as a convex problem. It is aimed to minimize the parameter size (indicated as the trace of precision matrix $\Sigma_{\mathrm{cm}}^{-1}$) and the violations of the margin ξ_i

$$\sum_{i=1}^{N}\xi_i + \epsilon\sum_{m=1}^{S}\sum_{k=1}^{K} trace\left(\Sigma_{\mathrm{m,k}}^{-1}\right) \quad (29)$$

subject to the constraints

$$\mathcal{F}(X_i,\widehat{W}_i|\lambda) - \mathcal{F}(X_i,W|\lambda) \geq \rho(W,\widehat{W}_i) - \xi_i,$$
$$\xi_i \geq 0, i = 1,\dots,N, \quad (30)$$
$$\Phi_{\mathrm{m,k}} \succcurlyeq 0, \mathrm{m} = 1,\dots,\mathrm{S}, \mathrm{k} = 1,\dots,\mathrm{M}.$$

for all $W \neq \widehat{W}_i$, where $\rho(W,\widehat{W}_i)$ is the MMR dependent on the competing word sequence, and $\Phi_{\mathrm{s,k}}$ is an augmented covariance matrix [16].

There are several strengths in LMT. First, while the number of the original constraints in (30) is huge, Sha et al. [16] proposed to effectively reduce the number of constraints by replacing them with a single stricter constraint (using the softmax function)

$$-\mathcal{T}(X_i,s_i^*) + log\sum_{s\neq s_i^*} e^{\rho(s,s_i^*)+\mathcal{F}(X_i,s)} \leq \xi_i \quad (31)$$

for each utterance, where the score of the oracle word sequence $\mathcal{T}(X_i,s_i^*)$ is approximated using the best path as in (5) and can be pre-computed using the initial model. Second, $\rho(W,\widehat{W}_i)$ and its approximation $\rho(s,s_i^*)$ are proportional to the number of disagreements (e.g., phoneme or state differences) between reference word sequence $\widehat{W}$ and other word sequences. Third, the training is carried out using the subgradient method (SGM) [16] which is applicable to large vocabulary tasks. Fourth, to alleviate the problem caused by the outliers, LMT used a rescaling approach so that the errors are bounded. Note, however, in formulating the optimization problem as a convex one, approximations need to be made. For example, the constraint that probabilities need to sum to one no longer holds, and the mixture components that maximize the likelihood (instead of the summation of the mixture components as in Eq. (1)) needs to be used, preselected, and considered as known in the training process.

Based on the work reported in [16], LMT with hidden states does not improve the performance over the MLE, while LMT with known states achieved 28.2% WER on the TIMIT corpus, which is 13.7% relative WER reduction over the MLE with the same model structure.

6. Summary and discussion

In this paper, we have reviewed four promising LMDT algorithms for estimating the parameters in HMMs. All these algorithms are aimed to optimize some form of the combined score of the empirical training-set error rate and the margin. Each of these existing algorithms has its own strengths and

weaknesses and none of them has surfaced as the obvious winner at this time. Table 1 summarizes these algorithms.

Table 1. Summary of large margin discriminative training algorithms for HMM

	LM-MCE	SME	LME, LRME	LMT
Convexity	No	No	Yes	Yes
Training Algorithm	GPD, EBW	GPD, EBW	GPD, SDP	SGM
Large Scale ASR Proven	Yes	Yes	No	No
Sensitive to Outliers	No	Yes	No	No
Error Types Used	String	String, Word, Phone	Likelihood Ratio	Likelihood Ratio

We believe that large-margin discriminative training is promising in improving the performance of large vocabulary ASR and we expect further WER reduction to be obtained using these techniques on large-scale tasks. To achieve this goal we suggest following future research directions.

First, design new theoretically sound training criteria that contain advantages of the existing ones and are free from their drawbacks.

Second, investigate better optimization algorithms. For example, can we formulate the training problem into a convex optimization (e.g., SDP) problem? Are there efficient online (instead of batch) algorithms for the problem?

Third, investigate, both theoretically and practically, the effects of large margin training on ASR performance and training complexity when very large amounts of training data (e.g., >10,000 hours) are made available.

Fourth, determine the percentage of data needed to achieve the same recognition accuracy using the LMDT techniques as that obtained by applying the traditional discriminative training on the entire data set. In other words, can we reduce the amount of data needed to train a good ASR model for languages with fewer resources? Can we reduce the overall training time for languages with large amount of training data?

7. References

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

[2] L. Deng and D. O'Shaughnessy, *Speech Processing - A Dynamic and Optimization-Oriented Approach*, Marcel Dekker Publishers, New York, NY, 2003.

[3] H. Jiang and X. Li, "Incorporating training errors for large margin HMMs under semi-definite programming framework", *Proc. ICASSP* 2007, vol. IV, pp. 629-632.

[4] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition", *IEEE Trans. Speech Audio Proc*., Vol. 5, May 1997.

[5] S. Kapadia, V. Valtchev, and S. J. Young, "MMI training for continuous phoneme recognition on the TIMIT database", *Proc. ICASSP* 1993, vol. 2, pp. 491-494.

[6] J. Li, M. Yuan, and C.-H. Lee, "Soft margin estimation of hidden Markov model parameters", *Proc. ICSLP* 2006, pp. 2422-2425.

[7] J. Li, S. M. Siniscalchi, and C.-H. Lee, "Approximate test risk minimization through soft margin estimation", *Proc. ICASSP* 2007, vol. IV, pp. 653-656.

[8] X. Li and H. Jiang, "A constrained joint optimization method for large margin HMM estimation", *Proc. IEEE ASRU Workshop*, 2005, pp. 151-156.

[9] X. Li, H. Jiang, and C. Liu, "Large margin HMMs for speech recognition", *Proc. ICASSP* 2005, pp. 513–516.

[10] X. Li and H. Jiang, "Solving large margin estimation of HMMs via semi-definite programming", *Proc. ICSLP* 2006, pp. 2414-2417.

[11] C. Liu, H. Jiang, and X. Li, "Discriminative training of CDHMMs for maximum relative separation margin", *Proc. ICASSP* 2005, vol. I, pp. 101-104.

[12] C. Liu, H. Jiang, and L. Rigazio, "Recent improvement on maximum relative margin estimation of HMMs for speech recognition", *Proc. ICASSP* 2006, vol. I, pp. 269-272.

[13] E. McDermott, T. Hazen, J. L. Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large vocabulary speech recognition using minimum classification error", *IEEE Trans. Speech and Audio Proc*, vol. 15, no. 1, 2007, pp. 203-223.

[14] E. McDermott and S. Katagiri, "A Parzen window based derivation of minimum classification error from the theoretical Bayes classification risk", *Proc. ICSLP*, 2002. Vol. 4, pp. 2465-2468.

[15] F. Sha and L. Saul, "Large margin Gaussian mixture modeling for phonetic classification and recognition", *Proc. ICASSP* 2006, Vol. 1, pp. 265-268.

[16] F. Sha, *Large margin training of acoustic models for speech recognition*, Ph.D. thesis, University of Pennsylvania, 2007.

[17] V. Vapnik, *The nature of Statistical learning theory*, 2nd edition, Springer-Verlag, New York, 1999.

[18] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition", *Computer Speech and Language*, vol. 16, 2002. pp. 25–47.

[19] D. Yu, L. Deng, X. He, and A. Acero, "Use of incrementally regulated discriminative margins in MCE training for speech recognition", *Proc. ICSLP* 2006, pp. 2418–2421.

[20] D. Yu, L. Deng, X. He, and A. Acero, "Large-margin minimum classification error training for large-scale speech recognition tasks", *Proc. ICASSP* 2007, vol. IV, pp. 1137-1140.