Large Margin Training of Continuous Density Hidden Markov Models

Fei Sha[1] and Lawrence K. Saul[2]

[1] Yahoo!
701 First Avenue
Mail stop: 1MC-08
Sunnyvale, CA 94089
feisha@yahoo-inc.com
[2] Department of Computer Science and Engineering
University of California, San Diego
9500 Gilman Drive, Mail Code 0404
La Jolla, CA 92093-0404
lsaul@ucsd.edu

Abstract. Continuous density hidden Markov models (CD-HMMs) are an essential component of modern systems for automatic speech recognition (ASR). These models assign probabilities to the sequences of acoustic feature vectors extracted by signal processing of speech waveforms. In this chapter, we investigate a new framework for parameter estimation in CD-HMMs. Our framework is inspired by recent parallel trends in the fields of ASR and machine learning. In ASR, significant improvements in performance have been obtained by discriminative training of acoustic models. In machine learning, significant improvements in performance have been obtained by discriminative training of large margin classifiers. Building on both these lines of work, we show how to train CD-HMMs by maximizing an appropriately defined margin between correct and incorrect decodings of speech waveforms. We start by defining an objective function over a transformed parameter space for CD-HMMs, then describe how it can be optimized efficiently by simple gradient-based methods. Within this framework, we obtain highly competitive results for phonetic recognition on the TIMIT speech corpus. We also compare our framework for large margin training to other popular frameworks for discriminative training of CD-HMMs.

1 Introduction

Most modern speech recognizers are built from continuous density hidden Markov models (CD-HMMs). The hidden states in these CD-HMMs are used to model different phonemes or sub-phonetic elements, while the observed outputs are used to model acoustic feature vectors. The accuracy of the speech recognition in CD-HMMs depends critically on the careful parameter estimation of their transition and emission probabilities.

The simplest method for parameter estimation in CD-HMMs is the Expectation-Maximization (EM) algorithm. The EM algorithm is based on maximizing the *joint likelihood* of observed feature vectors and label sequences. It is widely used due to

its simplicity and scalability to large data sets, which are common in ASR. However, this approach has the weakness that the model parameters of CD-HMMs are not optimized for sequential classification: in general, maximizing the joint likelihood does not minimize the phoneme or word error rates, which are more relevant metrics for ASR.

Noting this weakness, researchers in ASR have studied alternative frameworks for parameter estimation based on conditional maximum likelihood [1], minimum classification error [2] and maximum mutual information [3]. The learning algorithms in these frameworks optimize *discriminative* criteria that more closely track actual error rates. These algorithms do not enjoy the simple update rules and relatively fast convergence of the EM algorithm, which maximizes the joint likelihood of hidden state sequences and acoustic feature vectors. However, carefully and skillfully implemented, they lead to lower error rates [4–6].

Recently, in a new approach to discriminative acoustic modeling, we proposed a framework for large margin training of CD-HMMs. Our framework explicitly penalizes incorrect decodings by an amount proportional to the number of mislabeled hidden states. It also gives rise to a convex optimization over the parameter space of CD-HMMs, thus avoiding the problem of spurious local minima. Our framework builds on ideas from many previous studies in machine learning and ASR. It has similar motivation as recent frameworks for sequential classification in the machine learning community [7–9], but differs in its focus on the real-valued acoustic feature representations used in ASR. It has similar motivation as other discriminative paradigms in ASR [1–4, 6, 10, 11], but differs in its goal of margin maximization and its formulation of the learning problem as a convex optimization. The recent margin-based approaches of [12–16] are closest in terms of their goals, but entirely different in their mechanics.

In this chapter, we describe our framework for large margin training and present a systematic comparison to other leading frameworks for parameter estimation in CD-HMMs. We compare large margin training not only to maximum likelihood (ML) estimation, but also to popular discriminative methods based on conditional maximum likelihood (CML) [1, 3] and minimum classification error (MCE) [2]. We investigate salient differences between CML, MCE, and large margin training through carefully designed experiments on the TIMIT speech corpus [17]. In particular, we compare the results from multiple phonetic recognizers trained with different parameterizations, initial conditions, and learning algorithms. In all other aspects, though, these systems were held fixed: they employed exactly the same acoustic front end and model architectures (e.g., monophone CD-HMMs with full Gaussian covariance matrices). Our experimental results illuminate the significant factors that differentiate competing methods for discriminative training of CD-HMMs.

The paper is organized as follows. In section 2, we review CD-HMMs and current popular methods for parameter estimation. In section 3, we describe our framework for large margin training of CD-HMMs and give a brief survey of closely related work. In section 4, we compare the performance of phonetic recognizers trained in various different ways. Finally, in section 5, we review our main findings andf conclude with a brief discussion of future directions for research.

2 Background

CD-HMMs are used to specify a joint probability distribution over hidden state sequences $S = \{s_1, s_2, \ldots, s_T\}$ and observed output sequences $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$. The logarithm of this joint distribution is given by:

$$\log P(\boldsymbol{X}, S) = \sum_t \left[\log P(s_t|s_{t-1}) + \log P(\boldsymbol{x}_t|s_t)\right]. \tag{1}$$

For ASR, the hidden states s_t and observed outputs $\boldsymbol{x}_t$ denote phonetic labels and acoustic feature vectors, respectively, and the distributions $P(\boldsymbol{x}_t|s_t)$ are typically modeled by multivariate Gaussian mixture models (GMMs):

$$P(\boldsymbol{x}_t|s_t = j) = \sum_{m=1}^{M} \omega_{jm} \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}). \tag{2}$$

In eq. (2), we have used $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, while the constant M denotes the number of mixture components per GMM. The mixture weights ω_{jm} in eq. (2) are constrained to be nonnegative and normalized: $\sum_m \omega_{jm} = 1$ for all states j.

Let $\boldsymbol{\theta}$ denote all the model parameters including transition probabilities, mixture weights, mean vectors, and covariance matrices. The goal of parameter estimation in CD-HMMs is to compute the optimal $\boldsymbol{\theta}^*$ (with respect to a particular measure of optimality), given N pairs of observation and target label sequences $\{\boldsymbol{X}_n, Y_n\}_{n=1}^N$. In what follows, we review well-known frameworks for parameter estimation based on maximum likelihood (ML), conditional maximum likelihood (CML), and minimum classification error (MCE).

2.1 Maximum likelihood estimation

The simplest approach to parameter estimation in CD-HMMs maximizes the joint likelihood of output and label sequences. The corresponding estimator is given by

$$\boldsymbol{\theta}^{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_n \log P(\boldsymbol{X}_n, Y_n) \tag{3}$$

For transition probabilities, ML estimates in this setting are obtained from simple counts (assuming the training corpus provides phonetic label sequences). For GMM parameters, the EM algorithm provides iterative update rules that converge monotonically to local stationary points of the likelihood. The main attraction of the EM algorithm is that no free parameters need to be tuned for its convergence.

2.2 Conditional maximum likelihood

CD-HMMs provide transcriptions of unlabeled speech by inferring the hidden label sequence with the highest posterior probability: $Y = \arg\max_S P(S|\boldsymbol{X})$. The CML estimator in CD-HMMs directly attempts to maximize the probability that this inference

returns the correct transcription. Thus, it optimizes the conditional likelihood:

$$\boldsymbol{\theta}^{\mathrm{CML}} = \arg\max_{\boldsymbol{\theta}} \sum_n \log P(Y_n|\boldsymbol{X}_n). \tag{4}$$

In CML training, the parameters are adjusted to increase the likelihood gap between correct labelings Y_n and incorrect labelings S. This can be seen more explicitly by rewriting eq. (4) as:

$$\boldsymbol{\theta}^{\mathrm{CML}} = \arg\max_{\boldsymbol{\theta}} \left[\log P(\boldsymbol{X}_n, Y_n) - \log \sum_S P(\boldsymbol{X}_n, S) \right]. \tag{5}$$

The CML estimator in eq. (4) is closely related to the maximum mutual information (MMI) estimator [18, 6], given by:

$$\boldsymbol{\theta}^{\mathrm{MMI}} = \arg\max_{\boldsymbol{\theta}} \sum_n \log \frac{P(\boldsymbol{X}_n, Y_n)}{P(\boldsymbol{X}_n)P(Y_n)}. \tag{6}$$

Note that eqs. (4) and (6) yield identical estimators in the setting where the (language model) probabilities $P(Y_n)$ are held fixed.

2.3 Minimum classification error

MCE training is based on minimizing the number of sequence misclassifications. The number of such misclassifications is given by:

$$\mathcal{N}_{\mathrm{err}} = \sum_n \mathrm{sign}\left[-\log P(\boldsymbol{X}_n, Y_n) + \max_{S \neq Y_n} \log P(\boldsymbol{X}_n, S) \right] \tag{7}$$

where $\mathrm{sign}[z] = 1$ for $z > 0$ and $\mathrm{sign}[z] = 0$ for $z \leq 0$. To minimize eq. (7), the parameters must be adjusted to maintain a likelihood gap between the correct labeling and all competing labelings. Unlike CML training, however, the size of the gap in eq. (7) does not matter, as long as it is finite.

The nondifferentiability of the sign and $\max$ functions in eq. (7) makes it difficult to minimize the misclassification error directly. Thus, MCE training [5] adopts the surrogate cost function:

$$\mathcal{N}_{\mathrm{err}} \approx \sum_n \sigma\left(-\log P(\boldsymbol{X}_n, Y_n) + \log \left[\frac{1}{C} \sum_{S \neq Y_n} e^{\eta \log P(\boldsymbol{X}_n, S)} \right]^{\frac{1}{\eta}} \right), \tag{8}$$

where the sigmoid function $\sigma(z) = (1 + e^{-\alpha z})^{-1}$ replaces the sign function $\mathrm{sign}[z]$, and a softmax function (parameterized by η) replaces the original $\max$. The parameters α and η in this approximation must be set by heuristics. The sum in the second term is taken over the top C competing label sequences.

3 Large margin training

Recently, we proposed a new framework for discriminative training of CD-HMMS based on the idea of margin maximization [19]. Our framework has two salient features: (i) it attempts to separate the accumulated scores of correct versus incorrect label sequences by margins proportional to the number of mislabeled states [9]; (ii) the required optimization is convex, thus avoiding the pitfall of spurious local minima.

3.1 Discriminant function

We start by reviewing the discriminant functions in large margin CD-HMMs. These parameterized functions of observations $\boldsymbol{X}$ and states S take a form analogous to the log-probability in eq. (1). In particular, we define

$$\mathcal{D}(\boldsymbol{X}, S) = \sum_t \left[\lambda(s_{t-1}, s_t) + \rho(\boldsymbol{x}_t, s_t)\right] \tag{9}$$

in terms of state-state transition scores $\lambda(s_{t-1}, s_t)$ and state-output emission scores $\rho(\boldsymbol{x}_t, s_t)$. Unlike eq. (1), however, eq. (9) does not assume that the transition scores $\lambda(s_{t-1}, s_t)$ are derived from the logarithm of normalized probabilities. Likewise, the emission scores $\rho(\boldsymbol{x}_t, s_t)$ in eq. (9) are parameterized by sums of *unnormalized* Gaussian distributions:

$$\rho(\boldsymbol{x}_t, s_t = j) = \log \sum_m e^{-(\boldsymbol{x}_t - \boldsymbol{\mu}_{jm})^{\mathrm{T}} \boldsymbol{\Sigma}_{jm}^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_{jm}) - \theta_{jm}}, \tag{10}$$

where the nonnegative scalar parameter $\theta_{jm} \geq 0$ is entirely independent of $\boldsymbol{\Sigma}_{jm}$ (as opposed to being related to its log-determinant).

To obtain a convex optimization for large margin training, we further reparameterize the emission score in eq. (10). In particular, we express each mixture component's parameters $\{\boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}, \theta_{jm}\}$ as elements of the following matrix:

$$\boldsymbol{\Phi}_{jm} = \begin{bmatrix} \boldsymbol{\Sigma}_{jm}^{-1} & -\boldsymbol{\Sigma}_{jm}^{-1} \boldsymbol{\mu}_{jm} \\ -\boldsymbol{\mu}_{jm}^{\mathrm{T}} \boldsymbol{\Sigma}_{jm}^{-1} & \boldsymbol{\mu}_{jm}^{\mathrm{T}} \boldsymbol{\Sigma}_{jm}^{-1} \boldsymbol{\mu}_{jm} + \theta_{jm} \end{bmatrix}. \tag{11}$$

Our framework for large margin training optimizes the matrices $\boldsymbol{\Phi}_{jm}$, as opposed to the conventional GMM parameters $\{\boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}, \theta_{jm}\}$. Since the matrix $\boldsymbol{\Sigma}_{jm}$ is positive definite and the scalar θ_{jm} is nonnegative, we also require the matrix $\boldsymbol{\Phi}_{jm}$ to be positive semidefinite (as denoted by the constraint $\boldsymbol{\Phi}_{jm} \succ 0$). With this reparameterization, the emission score in eq. (10) can be written as:

$$\rho(\boldsymbol{x}_t, s_t = j) = \log \sum_m e^{-\boldsymbol{z}_t^{\mathrm{T}} \boldsymbol{\Phi}_{jm} \boldsymbol{z}_t} \quad \text{where} \quad \boldsymbol{z}_t = \begin{bmatrix} \boldsymbol{x}_t \\ 1 \end{bmatrix}. \tag{12}$$

Note that this score is convex in the elements of the matrices $\boldsymbol{\Phi}_{jm}$.

3.2 Margin constraints and Hamming distances

For large margin training of CD-HMMs, we seek parameters that separate the discriminant functions for correct and incorrect label sequences. Specifically, for each joint observation-label sequence $(\boldsymbol{X}_n, Y_n)$ in the training set, we seek parameters such that:

$$\mathcal{D}(\boldsymbol{X}_n, Y_n) - \mathcal{D}(\boldsymbol{X}_n, S) \;\geq\; \mathcal{H}(Y_n, S), \quad \forall\, S \neq Y_n \tag{13}$$

where $\mathcal{H}(Y_n, S)$ denotes the *Hamming distance* between the two label sequences [9]. Note how this constraint requires the log-likelihood gap between the target sequence Y_n and each incorrect decoding S to scale in proportion to the number of mislabeled states.

Eq. (13) actually specifies an exponentially large number of constraints, one for each alternative label sequence S. We can fold all these constraints into a single constraint by writing:

$$-\mathcal{D}(\boldsymbol{X}_n, Y_n) + \max_{S \neq Y_n} \{\mathcal{H}(Y_n, S) + \mathcal{D}(\boldsymbol{X}_n, S)\} \;\leq\; 0. \tag{14}$$

In the same spirit as the MCE derivation for eq. (8), we obtain a more tractable (i.e., differentiable) expression by replacing the $\max$ function in eq. (14) with a "softmax" upper bound:

$$-\mathcal{D}(\boldsymbol{X}_n, Y_n) + \log \sum_{S \neq Y_n} e^{\mathcal{H}(Y_n, S) + \mathcal{D}(\boldsymbol{X}_n, S)} \;\leq\; 0. \tag{15}$$

Note that the constraint in eq. (15) is stricter than the one in eq. (14); in particular, eq. (15) implies eq. (14). The exponential terms in eq. (15) can be summed efficiently using a modification of the standard forward-backward procedure.

3.3 Optimization

In general, it is not possible to find parameters that satisfy the large margin constraint in eq. (15) for all training sequences $\{\boldsymbol{X}_n, Y_n\}_{n=1}^N$. For such "infeasible" scenarios, we aim instead to minimize the total amount by which these constraints are violated. However, as a form of regularization, we also balance the total amount of violation (on the training sequences) against the scale of the GMM parameters. With this regularization, the overall cost function for large margin training takes the form:

$$\mathcal{L} \;=\; \gamma \sum_{cm} \text{trace}(\boldsymbol{\Phi}_{cm}) + \sum_n \left[-\mathcal{D}(\boldsymbol{X}_n, Y_n) + \log \sum_{S \neq Y_n} e^{\mathcal{H}(Y_n, S) + \mathcal{D}(\boldsymbol{X}_n, S)} \right]_+ . \tag{16}$$

The first term in the cost function regularizes the scale of the GMM parameters. In the second term, the "+" subscript in eq. (16) denotes the hinge function: $[z]_+ = z$ if $z > 0$ and $[z]_+ = 0$ if $z \leq 0$. The minimization of eq. (16) is performed subject to the positive semidefinite constraints $\boldsymbol{\Phi}_{jm} \succ 0$. We can further simplify the minimization by assuming that each emission score $\rho(\boldsymbol{x}_t, y_t)$ in the first term is dominated by the contribution from a single (pre-specified) Gaussian mixture component. In this case, the overall optimization is convex; see [19, 20] for further details. The gradients of this cost

function with respect to the GMM parameters $\boldsymbol{\Phi}_{cm}$ and transition parameters $\lambda(s, s')$ can be computed efficiently using dynamic programming, by a variant of the standard forward-backward procedure in HMMs [20].

It is worth emphasizing two crucial differences between this optimization and previous ones [1, 2, 6] for discriminative training of CD-HMMs for ASR. First, due to the reparameterization in eq. (11), the discriminant function $\mathcal{D}(\boldsymbol{X}_n, \boldsymbol{y}_n)$ and the softmax function are convex in the model parameters. Therefore, the optimization in eq. (15) can be cast as a convex optimization, avoiding spurious local minima [20]. Second, the optimization not only increases the log-likelihood gap between correct and incorrect state sequences, but also drives the gap to grow in proportion to the number of individually incorrect labels (which we believe leads to more robust generalization).

3.4 Related work

There have been several related efforts in ASR to incorporate ideas from large margin classification. Many studies have defined margins in terms of log-likelihood gaps between correct and incorrect decodings of speech. Previous approaches have also focused on balancing error rates versus model complexity. However, these approaches have differed from ours in important details. The approach in [12, 13] was based on maximizing the margin with respect to a *preselected subset* of training examples correctly classified by initial models. The parameters in this approach were re-estimated under constraints that they did not change too much from their initial settings. Later this approach was extended to include all training examples [16], but with the parameter estimation limited only to the mean vectors of GMMs. In [15], a large margin training criterion was defined in terms of averaged log-likelihood ratios over *mismatched frames*. This approach was motivated by earlier training criteria for minimum phone error training [21]. Finally, in other work inspired by large margin classification, Yu et al [14] extended MCE training by introducing and adapting nonzero offsets in the sigmoid approximation for the zero-one loss. By mimicking the effects of margin penalties, these offsets led to improved performance on a large vocabulary task in ASR.

Note that our work differs from all the above approaches by penalizing incorrect decodings in proportion to their Hamming distance from correct ones. In fact, as we discuss in section 4.2, it is possible to incorporate such penalties into more traditional frameworks for CML and MCE training. In very recent work, such an approach has been shown to reduce error rates on several tasks in large vocabulary ASR [22]. It is also possible to formulate margins in terms of string edit distances, as opposed to Hamming distances. Keshet et al [23] experimented with this approach in the setting of online parameter estimation.

4 Experimental results

We used the TIMIT speech corpus [17, 24, 25] to perform experiments in phonetic recognition. We followed standard practices in preparing the training, development, and test data. Our signal processing front-end computed 39-dimensional acoustic feature vectors from 13 mel-frequency cepstral coefficients and their first and second temporal

derivatives. In total, the training utterances gave rise to roughly 1.2 million frames, all of which were used in training.

We trained CD-HMMs for phonetic recognition in all of the previously described frameworks, including maximum likelihood (ML), conditional maximum likelihood (CML), minimum classification error (MCE), and margin maximization. We also experimented with several variants of these frameworks to explore the effects of different parameterizations, initializations, and cost functions.

In all the recognizers, the acoustic feature vectors were labeled by 48 phonetic classes, each represented by one state in a first-order CD-HMM. For each recognizer, we compared the phonetic state sequences obtained by Viterbi decoding to the "ground-truth" phonetic transcriptions provided by the TIMIT corpus. For the purpose of computing error rates, we followed standard conventions in mapping the 48 phonetic state labels down to 39 broader phone categories.

We computed two different types of phone error rates, one based on Hamming distance, the other based on edit distance. The former was computed simply from the percentage of mismatches at the level of individual frames. The latter was computed by aligning the Viterbi and ground truth transcriptions using dynamic programming [24], then summing the substitution, deletion, and insertion error rates from the alignment process. The "frame-based" phone error rate (based on Hamming distances) is more closely tracked by our objective function for large margin training, while the "string-based" phone error rate (based on edit distances) provides a more relevant metric for ASR. We mainly report the latter except in experiments where the frame-based error rate provides additional revealing context.

4.1 Large margin training

We trained two different types of large margin recognizers. The large margin recognizers in the first group were "low-cost" discriminative CD-HMMs whose GMMs were merely trained for frame-based classification. In particular, these GMMs were estimated by solving the (simpler) optimization for large margin training of isolated GMMs [26], then substituted into first-order CD-HMMs for sequence decoding. The large margin recognizers in the second group were fully trained for sequential classification. In particular, their CD-HMMs were estimated by solving the optimization in eq. (15) using multiple mixture components per state and adaptive transition parameters [19, 20, 27].

Tables 1 and 2 show the results of these experiments for CD-HMMs with different numbers of Gaussian mixture components per hidden state. For comparison, we also show the performance of baseline recognizers trained by ML estimation using the EM algorithm. For both types of phone error rates (frame-based and string-based), and across all model sizes, the best performance was consistently obtained by large margin CD-HMMs trained for sequential classification. Moreover, among the two different types of large margin recognizers, utterance-based training generally yielded significant improvement over frame-based training.

mixture components (per hidden state)	baseline ML (EM algorithm)	large margin (frame-based)	large margin (utterance-based)
1	45.2%	37.1%	29.5%
2	44.7%	36.0%	29.0%
4	42.2%	34.6%	28.4%
8	40.6%	33.8%	27.2%

Table 1. Frame-based phone error rates, computed from Hamming distance, of phonetic recognizers from differently trained CD-HMMs.

mixture components (per hidden state)	baseline ML (EM algorithm)	large margin (frame-based)	large margin (utterance-based)
1	40.1%	36.3%	31.2%
2	36.5%	33.5%	30.8%
4	34.7%	32.6%	29.8%
8	32.7%	31.0%	28.2%

Table 2. String-based phone error rates, computed from edit distance, of phonetic recognizers from differently trained CD-HMMs.

mixture components (per hidden state)	baseline ML (EM algorithm)	discriminative (CML)	discriminative (MCE)	discriminative (large margin)
1	40.1%	36.4%	35.6%	31.2%
2	36.5%	34.6%	34.5%	30.8%
4	34.7%	32.8%	32.4%	29.8%
8	32.7%	31.5%	30.9%	28.2%

Table 3. String-based phone error rates, computed from edit distance, of phonetic recognizers from differently trained CD-HMMs.

4.2 Comparison to CML and MCE

Next we compared large margin training to other popular frameworks for discriminative training. Table 3 shows the string-based phone error rates of different CD-HMMs trained by ML, CML, MCE, and margin maximization. As expected, all the discriminatively trained CD-HMMs yield significant improvements over the baseline CD-HMMs trained by ML. On this particular task, the results show that MCE does slightly better than CML, while the largest relative improvements are obtained by large margin training (by a factor of two or more). Using MMI on this task, Kapadia et al [11] reported larger relative reductions in error rates than we have observed for CML (though not better performance in absolute terms). It is difficult to compare our findings directly to theirs, however, since their ML and MMI recognizers used different front ends and numerical optimizations than those in our work.

mixture components (per hidden state)	normalized (CML)	unnormalized (CML)	normalized (MCE)	unnormalized (MCE)
1	36.4%	36.0%	35.6%	36.4%
2	34.6%	36.3%	34.5%	35.6%
4	32.8%	33.6%	32.4%	32.7%
8	31.5%	31.6%	30.9%	32.1%

Table 4. String-based phone error rates from discriminatively trained CD-HMMs with normalized versus unnormalized GMMs.

4.3 Other variants

What factors explain the better performance of CD-HMMs trained by margin maximization? Possible factors include: (i) the relaxation of Gaussian normalization constraints by the parameterization in eq. (11), yielding more flexible models, (ii) the convexity of the margin-based cost function in eq. (16), which ensures that its optimization (unlike those for CML and MCE) does not suffer from spurious local minima, and (iii) the closer tracking of phonetic error rates by the margin-based cost function, which penalizes incorrect decodings in direct proportion to their Hamming distance from the target label sequence. To determine which (if any) of these factors played a significant role, we conducted several experiments on the TIMIT corpus with variants of CML and MCE training.

First, we experimented with CML and MCE training procedures that did not enforce GMM normalization constraints. In these experiments, we optimized the usual objective functions for CML and MCE training, but parameterized the CD-HMMs in terms of the discriminant functions for large margin training in eq. (9). Table 4.3 compares phone error rates for CML and MCE training with normalized versus unnormalized GMMs. The table shows that lifting the normalization constraints generally leads to slightly worse performance, possibly due to overfitting. It seems that the extra degrees of freedom in unnormalized GMMs help to optimize the objective functions for CML and MCE training in ways that do not correlate with the actual phone error rate.

Next we experimented with different initial conditions for CML and MCE training. Because the optimizations in these frameworks involve highly nonlinear, non-convex objective functions, the results are susceptible to spurious local optima. To examine the severity of this problem, we experimented by re-initializing the CML and MCE training procedures with GMMs that had been discriminatively trained for segment-based phonetic classification [26]. These discriminatively trained GMMs provide a much better initialization than those trained by maximum likelihood estimation. Table 5 compares the results from different initializations. For both CML and MCE training, the differences in performance are quite significant: indeed, the improved results approach the performance of CD-HMMs trained by margin maximization. These results highlight a significant drawback of the non-convex optimizations in CML and MCE training– namely that the final results can be quite sensitive to initial conditions.

Finally, we experimented with variants of CML and MCE training that penalize incorrect decodings in proportion to their Hamming distances from the target label se-

mixtures components (per hidden state)	baseline (CML)	improved (CML)	baseline (MCE)	improved (MCE)
1	36.4%	32.6%	35.6%	32.9%
2	34.6%	31.7%	34.5%	31.3%
4	32.8%	31.2%	32.4%	31.1%
8	31.5%	28.9%	30.9%	29.0%

Table 5. String-based phone error rates from discriminatively trained CD-HMMs with differently initialized GMMs. The baseline GMMs were initialized by ML estimation. The improved GMMs were initialized by large margin training for segment-based phonetic classification.

quence. The Hamming distance penalties in large margin training put more emphasis on correcting egregiously mislabeled sequences. Our final experiments were designed to test whether similarly proportional penalties in CML and MCE training would lead to better performance. For CML training in these experiments, we maximized a reweighted version of the conditional likelihood:

$$\boldsymbol{\Theta}_{\mathcal{H}}^{\mathrm{CML}} = \arg\max_{\boldsymbol{\Theta}} \sum_n \log \frac{P(\boldsymbol{X}_n, Y_n)}{\sum_S e^{\mathcal{H}(Y_n,S)} P(\boldsymbol{X}_n, S)}. \tag{17}$$

The reweighting in eq. (17) penalizes incorrect decodings in proportion to their Hamming distance from the target label sequence, analogous to the cost function of eq. (16) for large margin training. For MCE training in these experiments, we applied the same intuition by incorporating the Hamming distance penalty into the surrogate cost function of eq. (8):

$$\mathcal{N}_{\mathrm{err}}^{\mathcal{H}} \approx \sum_n \sigma\left(-\log P(\boldsymbol{X}_n, Y_n) + \log\left[\frac{1}{C}\sum_{S\neq Y_n} e^{\eta[\mathcal{H}(Y_n,S)+\log P(\boldsymbol{X}_n,S)]}\right]^{\frac{1}{\eta}}\right). \tag{18}$$

Note that this approach differs from merely incorporating an adaptive scalar offset into the sigmoid transfer function. Such an offset has been shown to improve performance in large vocabularly ASR [14]. Though a positive scalar offset plays a similar role as a "margin", it does not explicitly penalize incorrect decodings in proportion to the number of mislabeled states. Table 6 shows the results from CML and MCE training with normal versus reweighted objective functions. Interestingly, the reweighting led to generally improved performance, but this positive effect diminished for larger models.

The experiments in this section were designed to examine the effects of relaxed normalization constraints, local versus global optima, and Hamming distance penalties in CML and MCE training. Overall, though better initializations and reweighted objective functions improved the results from CML and MCE training, none of the variants in this section ultimately matched the performance from large margin training. It is difficult to assess which factors contributed most strongly to this better performance. We suspect that *all* of them, working together, play an important role in the overall performance of large margin training.

mixtures components (per hidden state)	baseline (CML)	reweighted (CML)	baseline (MCE)	reweighted (MCE)
1	36.4%	33.6%	35.6%	33.3%
2	34.6%	32.8%	34.5%	32.3%
4	32.8%	32.8%	32.4%	31.5%
8	31.5%	31.0%	30.9%	30.8%

Table 6. String-based phone error rates from discriminatively trained CD-HMMs with different cost functions. The CD-HMMs in the third and fifth columns were trained by reweighting the penalties on incorrect decodings to grow in proportion to their Hamming distance from the target label sequence.

5 Conclusion

Discriminative learning of sequential models is an active area of research in both ASR [12, 4, 6] and machine learning [7–9]. In this chapter, we have described a particular framework for large margin training of CD-HMMs, which makes contributions to lines of work in both communities. In distinction to previous work in ASR, we have proposed a convex, margin-based cost function that penalizes incorrect decodings in proportion to their Hamming distance from the desired transcription. The use of the Hamming distance in this context is a crucial insight from earlier work [9] in the machine learning community, and it differs profoundly from merely penalizing the log-likelihood gap between incorrect and correct transcriptions, as commonly done in ASR. In distinction to previous work in machine learning, we have proposed a framework for sequential classification that naturally integrates with the infrastructure of modern speech recognizers. For real-valued observation sequences, we have shown how to train large margin HMMs via convex optimizations over their parameter space of positive semidefinite matrices. Using the softmax function, we have also proposed a novel way to monitor the exponentially many margin constraints that arise in sequential classification.

On the task of phonetic recognition, we compared our framework for large margin training to two other leading frameworks for discriminative training of CD-HMMs. In our experiments, CD-HMMs trained by margin maximization achieved significantly better performance than those trained by CML or MCE. Follow-up experiments suggested two possible reasons for this better performance: (i) the convexity of the optimization for large margin training and (ii) the penalizing of incorrect decodings in direct proportion to the number of mislabeled states. In future research, we are interested in applying large margin training to large vocabulary ASR, where both CML and MCE training have already demonstrated significant reductions in word error rates [18, 6, 5].

References

1. Nádas, A.: A decision-theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood. IEEE Transactions on Acoustics, Speech and Signal Processing **31**(4) (1983) 814–817

2. Juang, B.H., Katagiri, S.: Discriminative learning for minimum error classification. IEEE Transactions on Signal Processing **40**(12) (1992) 3043–3054
3. Bahl, L.R., Brown, P.F., de Souza, P.V., Mercer, R.L.: Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Tokyo (1986) 49–52
4. Roux, J.L., McDermott, E.: Optimization methods for discriminative training. In: Proceedings of Ninth European Conference on Speech Communication and Technology (EuroSpeech-05), Lisbon, Portgual (2005) 3341–3344
5. McDermott, E., Hazen, T.J., Roux, J.L., Nakamura, A., Katagiri, S.: Discriminative training for large vocabulary speech recognition using minimum classification error. IEEE Transactions on Speech and Audio Processing (2006)
6. Woodland, P.C., Povey, D.: Large scale discriminative training of hidden Markov models for speech recognition. Computer Speech and Language **16** (2002) 25–47
7. Altun, Y., Tsochantaridis, I., Hofmann, T.: Hidden Markov support vector machines. In Fawcett, T., Mishra, N., eds.: Proceedings of the Twentieth International Conference on Machine Learning (ICML-03), Washington, DC (2003) 3–10
8. Lafferty, J., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilisitc models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning (ICML-01), Morgan Kaufmann, San Francisco, CA (2001) 282–289
9. Taskar, B., Guestrin, C., Koller, D.: Max-margin Markov networks. In Thrun, S., Saul, L., Schölkopf, B., eds.: Advances in Neural Information Processing Systems (NIPS 16). MIT Press, Cambridge, MA (2004) 25–32
10. Gopalakrishnan, P.S., Kanevsky, D., Nádas, A., Nahamoo, D.: An inequality for rational functions with applications to some statistical estimation problems. IEEE Transactions on Information Theory **37**(1) (1991) 107—113
11. Kapadia, S., Valtchev, V., Young, S.: MMI training for continuous phoneme recognition on the TIMIT database. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-93). Volume 2., Minneapolis, MN (1993) 491–494
12. Li, X., Jiang, H., Liu, C.: Large margin HMMs for speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-05), Philadelphia, PA (2005) 513–516
13. Jiang, H., Li, X., Liu, C.: Large margin hidden Markov models for speech recognition. IEEE Transactions on Audio, Speech and Language Processing **14**(5) (2006) 1584–1595
14. Yu, D., Deng, L., He, X., Acero, A.: Use of incrementally regulated discriminative margins in MCE training for speech recognition. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP-06), Pittsburgh, PA (2006) 2418–2421
15. Li, J., Yuan, M., Lee, C.: Approximate test risk bound minimization through soft margin estimation. IEEE Transactions on Speech, Audio and Language Processing **15**(8) (2007) 2392–2404
16. Jiang, H., Li, X.: Incorporating training errors for large margin HMMs under semi-definite programming framework. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-07). Volume 4., Honolulu, HI (2007) 629–632
17. Lamel, L.F., Kassel, R.H., Seneff, S.: Speech database development: design and analsysis of the acoustic-phonetic corpus. In Baumann, L.S., ed.: Proceedings of the DARPA Speech Recognition Workshop. (1986) 100–109
18. Valtchev, V., Odell, J.J., Woodland, P.C., Young, S.J.: MMIE training of large vocabulary recognition systems. Speech Communication **22** (1997) 303–314
19. Sha, F., Saul, L.K.: Large margin hidden Markov models for automatic speech recognition. In Schölkopf, B., Platt, J., Hofmann, T., eds.: Advances in Neural Information Processing Systems 19, Cambridge, MA, MIT Press (2007) 1249–1256

20. Sha, F.: Large margin training of acoustic models for speech recognition. PhD thesis, University of Pennsylvania, Philadelphia, PA (2007)
21. Povey, D., Woodland, P.: Minimum phone error and i-smoothing for improved discriminative training. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-02), Orlando, FL (2002) 105–108
22. Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., Visweswariah, K.: Boosted MMI for model and feature-space discriminative training. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-08), Las Vegas, NV (2008)
23. Keshet, J., Shalev-Shwartz, S., Bengio, S., Singer, Y., Chazan, D.: Discriminative kernel-based phoneme sequence recognition. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP-06), Pittsburgh, PA (2006)
24. Lee, K.F., Hon, H.W.: Speaker-independent phone recognition using hidden Markov models. IEEE Transactions on Acoustics, Speech, and Signal Processing **37**(11) (1988) 1641–1648
25. Robinson, T.: An application of recurrent nets to phone probability estimation. IEEE Transactions on Neural Networks **5**(2) (1994) 298–305
26. Sha, F., Saul, L.K.: Large margin Gaussian mixture modeling for phonetic classification and recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-06), Toulouse, France (2006) 265–268
27. Sha, F., Saul, L.K.: Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-07), Honolulu, HI (2007) 313–316