Discriminative Training of Tied-Mixture HMM by Deterministic Annealing

Liang Gu, Jayanth Nayak and *Kenneth Rose*

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106, USA
Email: {liang, jayanth, rose}@scl.ece.ucsb.edu

ABSTRACT

A deterministic annealing algorithm for the design of tied-mixture HMM recognizers is proposed, which reduces the training sensitivity to parameter initialization, automatically smoothes the classification error cost function to allow gradient-based optimization, and seeks better solutions than known techniques. The new approach introduces randomness into the classification rule during the training process, and minimizes the expected error rate while controlling the level of randomness via a constraint on the Shannon entropy. As the entropy constraint is gradually relaxed, the effective cost function converges to the classification error rate and the system becomes a hard (non-random) recognizer. Experiments show that the proposed method outperforms design by maximum likelihood re-estimation and by generalized probabilistic descent.

1. INTRODUCTION

Tied-mixture Hidden Markov Modeling (TMHMM) [1][2] is recognized as a useful complexity reduction technique for robust speech recognition, mainly due to its ability to maintain modeling accuracy of large-mixture probability density functions (pdfs) at moderate complexity, by enforcing pdf sharing. However, the large number of additional mixing coefficients that TMHMM introduces, along with the universal set (or codebook) of pdfs, presents a new design challenge. While the conventional expectation-maximization (EM) algorithms [1] seem satisfactory in speed and performance, they tend to converge to suboptimal solutions and strongly sensitive to parameter initialization [2]. This motivates the search for more robust training methods to exploit the true potential of TMHMM.

Another important concern is the training criterion. While the natural speech recognition objective is minimum classification error (MCE), HMM design has been traditionally approached via maximum likelihood (ML) which, although mismatched with MCE and hence suboptimal, circumvents the difficulties due to the piecewise-constant MCE cost function that resists direct attacks by gradient-based optimization methods. Recently, there appeared several new MCE techniques, notably the generalized probabilistic descent (GPD) approach (see review in [3]), which smooth the classification error cost function and jointly optimize all HMM parameters via gradient descent. These methods target the true design cost and thereby offer the potential for significant performance gains over ML. However, the smoothed MCE cost surface is riddled with shallow local minima that easily trap local descent methods, and may substantially compromise performance.

A natural approach to tackle the above difficulties is to introduce powerful optimization tools into the training procedure. The deterministic annealing (DA) algorithm has been shown to be an effective optimization tool for similar tasks [4]. Derived from fundamental principles of statistical physics and information theory, DA was first proposed for clustering and related problems [5][6], and later applied to pattern classifiers [7], source coding systems [8], regression functions [9], etc. Most recently, DA has been successfully applied in the design of discrete observation HMM (DHMM) [10][11] and continuous density HMM (CHMM) recognizers [12], and was shown to substantially outperform both ML-based EM algorithm and MCE-based GPD algorithm.

In this paper, we propose a new DA-based training algorithm to design recognizers based on TMHMM, which can substantially reduce the training sensitivity to parameter initialization, automatically smooth the MCE cost function, and find better solutions than existing techniques. To achieve these goals, the standard DA procedure is adjusted to the design of TMHMM, particularly to account for the mixing coefficients and universal pdfs. Fast forward-backward algorithm is further developed to reduce the computational complexity. Experimental results on the E-set show that the MCE-based DA algorithm outperforms both the ML-based EM and MCE-based GPD algorithms, and that the sensitivity to parameter initialization is substantially reduced by DA.

2. DETERMINISTIC ANNEALING FOR DISCRIMINATIVE TMHMM TRAINING

A. Deterministic vs. stochastic annealing

Annealing is a process where a physical system is gradually cooled, starting at a sufficiently high temperature, while maintaining the system in thermal equilibrium at all intermediate temperatures. Recent years have seen several powerful optimization algorithms that exploit the analogy between optimization and the physical process of annealing, including the popular method of stochastic annealing (SA) or simulated annealing [13]. SA simulates the random evolution of a physical system and reaches equilibrium as the steady-state distribution over states of a corresponding Markov chain. If the annealing schedule is sufficiently slow, SA can be shown to asymptotically converge in probability to the set of globally optimal solutions.

This work was supported in part by the National Science Foundation under grant no. IIS-9978001, the University of California MICRO Program, Conexant Systems, Inc., Fujitsu Laboratories of America, Inc, Lernout & Hauspie Speech Products, Lucent Technologies, Inc., and Qualcomm, Inc.

DA is also based on the annealing process, albeit in a much different way. Instead of simulating the exact stochastic evolution of the system, DA efficiently employs expectation. Specifically, it determines the effective distribution over the states of the system at each temperature and optimizes the expected value of the cost function (the free energy in the physical analogy). Thus, DA does not generate a stochastic process that evolves via numerous "moves" (modification of system parameters) per temperature in order to reach thermal equilibrium as in SA. Rather, it directly optimizes the free energy for each specific temperature, which is in fact the thermodynamic quantity minimized stochastically by SA. DA may hence be viewed as a deterministic relative of SA. Although DA offers no guarantee of finding the global optimum, it does inherit from the annealing process an ability to effectively avoid many local minima.

B. Problem statement

Let the HMM-based speech recognizer be trained from the labeled training set $T = \{(\bar{x}_1(t), c_1), (\bar{x}_2(t), c_2), \cdots, (\bar{x}_L(t), c_L)\}$, where $\bar{x}_i(t)$ represents l_i p-dimensional sequential feature vectors extracted from a speech sample of class c_i, which belongs to the finite-size dictionary $C = \{c_1, c_2, \cdots, c_N\}$. Without loss of generality, we assume that the recognizer has *N* models, $\{H_j, j = 1,2,\cdots, N\}$, one per word in dictionary *C*. Each HMM H_j is fully specified by the parameter set $\Lambda_j = (A_j, B_j, \Pi_j)$, where A_j specifies the state transition probabilities, B_j contains the state-conditional emission distributions, and Π_j specifies the initial state probabilities. For concreteness we further assume that the recognizer employs the "best path" discriminant. The approach can be similarly derived for the case where the discriminant is computed by likelihood averaging over all paths in the HMM.

For sample $\bar{x}_i(t)$ and model H_j, the normalized logarithm of the joint probability ("path score") of observation $\bar{x}_i(t)$ for a sequence of states $s \equiv (s(1), s(2), \cdots, s(l_i))$ in the trellis of H_j is defined as

$$l(x_i, s, H_j) = \frac{1}{l_i}\left\{\begin{aligned}&\log \Pi_j[s(1)] + \sum_{t=1}^{l_i - 1} \log A_j[s(t), s(t+1)] \\ &+ \sum_{t=1}^{l_i} \log B_j[s(t), x_i(t)]\end{aligned}\right\}.$$

The discriminant is computed by maximizing the score over all paths:

$$d_j(x_i) = \max_{s \in S_{l_i}(H_j)} l(x_i, s, H_j),$$

where $S_{l_i}(H_j)$ is the set of all state sequences of length l_i in the trellis of H_j.

The traditional "best path" classification rule is

$$C(x_i) = \arg\max_j d_j(x_i).$$

An MCE design method optimizes the HMM parameters $\{\Lambda_j\}$ so as to minimize the misclassification rate measured over the training set, i.e.,

$$\min_{\{\Lambda_j\}} \left\{ P_e = 1 - \frac{1}{L} \sum_{i=1}^{L} d(C(x_i), c_i) \right\}, \quad (1)$$

where d is the Kronecker delta function:

$$d(u, v) = \begin{cases} 1, & if\ u = v \\ 0, & otherwise. \end{cases}$$

An immediate difficulty with MCE-based training is due to the piecewise constant nature of (1), which does not lend itself to gradient-based optimization. The ML approach circumvents this problem by substituting the true cost function with a sub-optimal design objective. GPD and other MCE algorithms smooth the MCE cost function to allow descent-based optimization, but still suffer from the poor local optima problem.

C. Discriminative HMM training by deterministic annealing

DA offers means to avoid many poor local optima while implementing a theoretically motivated form of cost smoothing. To achieve this goal, three fundamental principles are employed: a) Introduce randomness in the recognition rule during the training process; b) Minimize the expected error rate of the random recognizer while controlling the level of randomness via a constraint on the Shannon entropy; and c) Gradually relax the entropy constraint so that the effective cost converges to standard MCE at the limit of zero entropy (non-random classification).

Principle I: *The randomized "soft" recognizer*

Randomness is introduced into the recognition rule during the training process, and the "best-path" ("hard") recognizer is replaced by a "soft" recognizer. Instead of assigning training sample x_i to a unique winning state sequence in the trellis of model H_j, the randomized rule associates it with every state sequence *s* with probability $p(s, H_j | x_i)$, which is the Gibbs distribution

$$p(s, H_j | x_i) = \frac{e^{g \cdot l(x_i, H_j, s)}}{\sum_{k=1}^{N} \sum_{s \in H_k} e^{g \cdot l(x_i, H_k, s)}},$$

where g is a parameter that controls the fuzziness of the distribution. When $g = 0$, the distribution over paths is uniform and the recognizer is extremely random (i.e. the recognition rate for training set is 1/N). For a higher value of g, the randomized recognizer assign the paths of higher log probabilities with higher path scores. In the extreme case of

$g \to \infty$, the random recognizer becomes a non-random "best path" recognizer. Note that the random recognizer is only used during training, and the actual resulting system is of course non-random.

Principle II: *Minimization of the effective cost function*

While the recognizer is random we consider the expected error rate criterion given by

$$\langle P_e \rangle = 1 - \frac{1}{L} \sum_{i=1}^{L} \sum_{j=1}^{N} \left[d\left(C(x_i), c_i\right) \cdot \sum_{s \in H_j} p\left(s, H_j \middle| x_i\right) \right] \quad (2)$$

Note that when $g \to \infty$, $\langle P_e \rangle$ becomes the standard error rate P_e of (1). As the piecewise function P_e was also smoothed by randomization, and direct gradient-based optimization of (2) is possible though highly susceptible to shallow local minimum traps. Instead, we propose to minimize $\langle P_e \rangle$ subject to a constraint on the level of randomness that will enable us to introduce annealing as will be discussed below. The randomness is measured by the (conditional) Shannon entropy

$$H = -\frac{1}{L} \sum_{i} \sum_{j} \sum_{s \in S_{l_i}(H_j)} P\left(s, H_j \middle| x_i\right) \log P\left(s, H_j \middle| x_i\right) \quad (3)$$

The constrained optimization problem is equivalently expressed as the minimization of the unconstrained Lagrangian cost function:

$$\min_{\{\Lambda_j\}, g} \left\{ F \equiv \langle P_e \rangle - TH \right\}, \quad (4)$$

where F is the effective cost function and T is the Lagrange multiplier which is referred to as the "temperature" to allude the interesting analogy to statistical physics, while F is equivalent to the Helmholtz free energy of a thermodynamic system. The optimization of (4) is analogous to achieving thermal equilibrium at the given T. When $T \to 0$, the optimization reduces to the unconstrained minimization of $\langle P_e \rangle$. The gradual reduction of is important to avoid shallow local minima on the cost function surface.

Principle III: *Annealing process*

The DA-based HMM training process is described in the flow diagram of Figure 1. The system starts at high temperature T and high randomness (low g), and gradually decreases T in analogy to physical annealing. At each intermediate temperature, F is minimized via gradient-based optimization of the HMM parameters and g. As the temperature decreases, g naturally increases and reduces both the randomness of recognizer (i.e., the Shannon entropy *H)*. Hence, the distribution becomes more discriminating and, ultimately, only the most likely path is assigned a non-zero probability. The resulting recognizer is therefore a normal, non-random "best-path" recognizer.

In practice, it is convenient to accelerate the final elimination of randomness via a quenching phase as shown in Figure 1: During

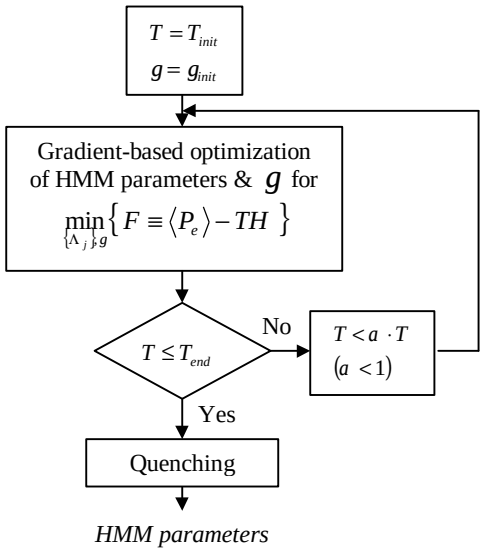

Figure 1. DA procedure for HMMM training

the main DA procedure, g is upperbound by $g_{\max}$. When T is sufficiently small, we optimize the HMM parameters while increasing g gradually to a high value, and finally convert the random recognizer into a non-random "best-path" recognizer.

D. The DA update equations

From (2), (3) and (4), we have

$$F \equiv \langle P_e \rangle - TH = 1 + \frac{1}{L} \sum_{i} \sum_{j} \sum_{s \in S_{l_i}(H_j)} p\left(s, H_j \middle| x_i\right) \cdot \left\{ T \left[g \cdot l\left(x_i, s, H_j\right) - \log \left(\sum_{j'} \sum_{s' \in S_{l_i}(H_{j'})} e^{g \cdot l\left(x_i, s', H_{j'}\right)} \right) \right] - d\left(j, c_i\right) \right\}.$$

The gradient descent algorithm utilizes the following partial derivatives:

$$\frac{\partial F}{\partial l_k} = \frac{g}{L} \sum_{i} \sum_{j} \sum_{s \in S_{l_i}(H_j)} p\left(s, H_j \middle| x_i\right) \cdot \left[\frac{\partial l\left(x_i, s, H_j\right)}{\partial l_k} - \left\langle \frac{\partial l(x_i)}{\partial l_k} \right\rangle \right] \cdot f\left(x_i, s, H_j\right)$$

$$\frac{\partial F}{\partial g} = \frac{1}{L} \sum_{i} \sum_{j} \sum_{s \in S_{l_i}(H_j)} p\left(s, H_j \middle| x_i\right) \cdot \left[l\left(x_i, s, H_j\right) - \langle l(x_i) \rangle \right] \cdot f\left(x_i, s, H_j\right)$$

where

$$f\left(x_i, s, H_j\right) = Tg \cdot l\left(x_i, s, H_j\right) - d\left(j, c_i\right)$$

$$\left\langle \frac{\partial l(x_i)}{\partial l_k} \right\rangle = \sum_{j} \sum_{s \in S_{l_i}(H_j)} p\left(s, H_k \middle| x_i\right) \cdot \frac{\partial l\left(x_i, s, H_k\right)}{\partial l_k}$$

$$\langle l(x_i) \rangle = \sum_{j} \sum_{s \in S_{l_i}(H_j)} p\left(s, H_k \middle| x_i\right) \cdot l\left(x_i, s, H_k\right)$$

In practice, we optimize g by line-search. While the above allows for easy interpretation of the update rule, it is convenient for implementation to rewrite the derivative with respect to TMHMM parameters as

$$\frac{\partial F}{\partial l_k} = \frac{g}{L}\sum_i \{Tg \cdot I_1(x_i,k) + I_2(x_i,k)[I_4(x_i,c_i) - d(c_i,k) - Tg \cdot I_3(x_i)]\}$$

where

$$I_1(x_i,k) = \sum_j \sum_{s \in S_{l_i}(H_j)} p(s,H_j|x_i) \cdot \frac{\partial l(x_i,s,H_j)}{\partial l_k} \cdot l(x_i,s,H_j)$$

$$I_2(x_i,k) = \sum_j \sum_{s \in S_{l_i}(H_j)} p(s,H_j|x_i) \cdot \frac{\partial l(x_i,s,H_j)}{\partial l_k}$$

$$I_3(x_i) = \sum_j \sum_{s \in S_{l_i}(H_j)} p(s,H_j|x_i) \cdot l(x_i,s,H_j)$$

$$I_4(x_i,c_i) = \sum_{s \in S_{l_i}(H_{c_i})} p(s,c_i|x_i).$$

We have developed an efficient forward-backward algorithm similar to [11] to compute the above four sets of variables but will not detail it here for lack of space.

3. EXPERIMENTAL RESULTS

To test the performance of the proposed DA algorithm for TMHMM training, experiments were carried out on the E-set speech database obtained from OGI. The recognition task is to distinguish between nine confusable English letters {b, c, d, e, g, p, t, v, z}. The database was generated by 150 speakers (75 male and 75 female) and includes one utterance per speaker. Of the 150 speakers, 60 male and 60 female speakers were selected at random for training, and the remaining 30 speakers were set aside for the test set.

In our experiments, 12-dimension MFCC parameters were used as the speech features. The analysis frame width is 30ms, the frame step is 10ms, and a Hamming Window is employed. The results are summarized in Tables 1 and 2. Table 1 compares the performance of various TMHMM training methods at three states per HMM. The results demonstrate consistent performance improvement from the standard EM algorithm (ML criterion) through GPD (MCE criterion) to the proposed DA algorithm (MCE). Table 2 illustrates that the sensitivity to initialization has been greatly reduced by DA compared with the standard EM algorithm.

Recognition Rate	ML-EM	MCE-GPD	MCE-DA
Train Set	62.1%	70.2%	75.5%
Test Set	57.5 %	60.1%	63.9 %

Table 1. Performance comparison of TMHMM training methods at 3 states per HMM

Training method	Initialized From DHMM	Initialized from K-means	Initialized from CHMM
ML-EM	52.7 %	55.4%	57.5 %
MCE-DA	63.8%	63.8%	63.9%

Table 2. Test set recognition rate for different initializations

4. CONCLUSION

TMHMM training has long been a challenging task due to its complex structure (relative to standard CHMM). In this paper, a deterministic annealing (DA) algorithm was proposed for TMHMM training which substantially reduces the training sensitivity to parameter initialization, automatically smoothes the MCE cost function, and finds better solutions than GPD. Preliminary experiments on the E-set show that DA does offer improvement over ML-EM and MCE-GPD. Future work will expand on the potential of DA and in particular will address the problem of Gaussian selection.

5. REFERENCES

[1] J. R. Bellegarda and D. Nahamoo, "Tied mixture continuous parameter modeling for speech recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 2033-2045, vol. 38, Dec. 1990.

[2] X. D. Huang, "Phoneme classification using semicontinuous hidden Markov models", *IEEE Trans. Signal Processing*, vol. 40, pp. 1062-1067, May 1992

[3] S. Katagiri, B. H. Juang, and C. H. Lee, "Pattern recognition using a family of design algorithms based upon the generalized probabilistic decent method", *IEEE Proceedings*, vol. 86, no. 11, pp.2345-3375, 1998.

[4] K. Rose, "Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems", *IEEE Proceedings*, vol. 86, pp.2210-2239, 1998.

[5] K. Rose, E. Gurewitz, and G. C. Fox, "Vector quantization by deterministic annealing", *IEEE Trans. on Information Theory*, vol. 38, pp.1249-1258, 1992.

[6] K. Rose, E. Gurewitz, and G. C. Fox, "Constrained clustering as an optimization method", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, pp.785-794, 1993.

[7] D. Miller, A. V. Rao, K. Rose, and A. Gersho, "A global optimization techniques for statistical classifier design", *IEEE Trans. on Signal Processing*, vol. 44, no. 12, 1996.

[8] A. V. Rao, D. Miller, K. Rose, and A. Gersho, "A generalized VQ method for combined compression and estimation", *Proc. ICASSP'1996*, pp.2032-2035.

[9] A. V. Rao, D. Miller, K. Rose, and A. Gersho, "Mixture of experts regression modeling by deterministic annealing", *IEEE Trans. on Signal Processing*, vol. 45, no. 11, pp.2811-2820, 1997.

[10] A. V. Rao, K. Rose, and A. Gersho, "Design of robust HMM speech recognizers using deterministic annealing", *Proc. IEEE ASRU'97*, pp.466-473, 1997.

[11] A. V. Rao and K. Rose, "Deterministically annealed design of hidden Markov model speech recognizers", to appear in *IEEE Trans. on Speech and Audio Processing*.

[12] C. Gelin-Huet, K. Rose and A. V. Rao, "The deterministic annealing approach for discriminative continuous HMM design", *Proc. EuroSpeech'99*, 1999.

[13] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, "Optimization by simulated annealing", *Science*, vol. 220, pp.671-680, 1983.