A DETERMINISTIC ANNEALING APPROACH TO DISCRIMINATIVE HIDDEN MARKOV MODEL DESIGN *

Ajit Rao, Kenneth Rose, and Allen Gersho
Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106.

Abstract

We present the problem of designing a classifier system based on hidden Markov models (HMMs) from a labeled training set with the objective of minimizing the rate of mis-classification. The traditional design approach divides the training set into subsets of identically labeled training vectors and independently designs the HMM corresponding to each subset of the training data using a maximum likelihood criterion. However, this approach does not achieve the minimum mis-classification objective. To design the globally optimal recognizer, all the HMMs must be jointly optimized to minimize the number of mis-classified training patterns. This is a difficult design problem which we attack using the technique of deterministic annealing (DA). In the DA approach, we introduce randomness in the classification rule and minimize the expected mis-classification rate of the random classifier while controlling the level of randomness in its decision via a constraint on the Shannon entropy. The effective cost function is smooth and converges to the mis-classification cost at the limit of zero entropy (non-random classification rule). The DA approach can be implemented via an efficient forward-backward algorithm for recomputing the model parameters.

*This work was supported in part by the National Science Foundation under grant no. NCR-9314335, the University of California MICRO program, ACT Networks, Inc., Advanced Computer Communications, Cisco Systems, Inc., DSP Group, Inc., DSP Software Engineering, Inc., Fujitsu Laboratories of America, Inc., General Electric Company, Hughes Electronics Corp., Intel Corp., Nokia Mobile Phones, Qualcomm, Inc., Rockwell International Corp., and Texas Instruments, Inc.,

0-7803-4256-9/97/$10.00 ©1997 IEEE

This algorithm significantly outperforms the standard maximum likelihood algorithm for a moderate increase in design complexity.

1 Introduction

The hidden Markov model (HMM) is commonly used as a stochastic model for time sequences. HMMs were originally applied within main-stream statistics, but the discovery of their applicability to modeling speech utterances [3, 6] has led to extensive research activity in HMMs over the last three decades. An overwhelming number of conventional speech recognition systems are based on the use of the HMM to model various speech utterances within the context of traditional discriminant-based pattern classification.

In this paper, we address the problem of recognition of time sequences modeled by HMMs. It is formally defined as the design of a recognizer based on a labeled training set (i.e., supervised learning). This problem has been extensively treated in the speech recognition literature. The most commonly used approach is to divide the training set into subsets of identically labeled training vectors and independently design HMMs for each subset of training data via maximum likelihood estimation of model parameters. After design, the system is used for recognizing new sequences through competition between the designed HMMs. The input sequence is declared to belong to the winner (the most likely model).

The starting point of our work is the realization that the above recognition problem is fundamentally a pattern classification problem. Further, the quality of the recognizer is most appropriately measured by its rate of classification error. This leads to two major observations: First, the globally optimal recognizer must be designed through *joint optimization* of all models. It is important to emphasize that the ultimate objective is not to model the sequences belonging to each class as accurately as possible, but rather, to distinguish between the classes while making as few errors as possible. As classification is performed by competition between models, it is clear that we must optimize all the model parameters simultaneously to minimize classification errors.

This also connects to the second observation, namely, that maximum likelihood is a mismatched cost for optimizing the classifier. The direct measure of success is simply the empirical rate of correct classification. It should be noted in passing that the Bayesian classifier which is optimal in the sense of minimum classification error, is a close relative of the maximum likelihood approach above. However its success depends on the availability of the precise probability distributions, including the assumption that the model structure is in complete agreement with the source. If one has only access to a reasonably short training set, the performance of maximum likelihood

may differ significantly from that of minimum classification error, as will be demonstrated in this work. We note that the shortcomings of the maximum likelihood method have been previously recognized (*e.g.* [1, 2, 4, 7]) and joint optimization approaches have been suggested.

There are several important difficulties in approaching the design problem directly, that is, by joint optimization of all model parameters so as to minimize the rate of classification error. One difficulty is that unlike maximum likelihood, this cost function is piecewise constant and all gradients with respect to parameters vanish almost everywhere (an infinitesimal change in parameter values will not change the classification of any sequence in the training set). Thus, one cannot simply use a gradient based optimization method. An important approach to address this problem appeared in [7] where the cost surface was smoothed to allow the application of gradient methods (A few weeks ago, a paper appeared [5], where this method was extended to HMM classification.). Another important difficulty is that even if the cost surface is smoothed, the optimization process tends to suffer from numerous shallow local minima that riddle this complex cost surface. Finally, one must keep in mind the difficulties associated with the computational complexity of such joint optimization.

The main contribution of this paper is a novel method for designing HMM-based recognizers. The new method is based on the deterministic annealing approach to clustering [14, 13] and in particular to its recent extension to classification [8]. By introducing randomness that is controlled by imposing the level of Shannon entropy, we obtain an effective cost function that is smooth and converges to the original classification error cost at the limit of zero entropy. Further, this process is analogous to physical annealing and hence has the capability to avoid many shallow minima that trap standard local optimization methods. It is also important to note that unlike the stochastic procedure of simulated annealing, the process here is deterministic and all randomization is taken into account by taking the expectation of the various quantities. Another important result is the development of a forward-backward algorithm (similar to Baum-Welch re-optimization) for recomputing the parameters of all models in our joint optimization framework. (Note that here we do not use maximum likelihood as our ultimate objective). This algorithm is instrumental in keeping the computational complexity manageable. The approach is shown to substantially outperform the standard maximum likelihood method at the cost of moderate increase in design complexity with respect to separate design of HMM per class.

2 The HMM classifier and its design

We address the supervised learning problem of designing a recognition system from a labeled training set, $\mathcal{T} \equiv \{(\mathbf{y}_1, c_1), (\mathbf{y}_2, c_2), ..(\mathbf{y}_N, c_N)\}$. Each *train-*

ing pattern, $\mathbf{y}_i$, is a vector of l_i observations, $\mathbf{y}_i = (\mathbf{y}_i(1), \mathbf{y}_i(2), \cdots \mathbf{y}_i(l_i))$. Further, each observation, $\mathbf{y}_i(t)$, is a discrete quantity, *i.e.* $\mathbf{y}_i(t) \in \mathcal{A} \equiv \{1, 2, \cdots, K\}$. Despite this restriction to the case of discrete observations, we note that the design methods can be easily be extended to handle continuous valued observations also. The training pattern, $\mathbf{y}_i$, belongs to class, c_i, which may be one of M classes, *i.e.* $c_i \in \mathcal{C} \equiv \{1, 2, \cdots M\}$.

The HMM recognition system consists of a set of hidden Markov models, $\{H_j,\ j = 1, 2, \cdots, M\}$, one per class index. The model, H_j has S_j states and is fully specified by the parameter set $\Lambda_j \equiv (A_j, B_j, \Pi_j)$, where following the usual convention, A_j is the $(S_j \times S_j)$ state transition probability matrix, B_j is the $(S_j \times K)$ emission probability matrix and Π_j is the (length S_j) initial state probability vector.

The classifier works as follows : Given a training pattern, $\mathbf{y}_i$, for each HMM, H_j, and for each sequence (length l_i) of states, $\mathbf{s} \equiv (s(1), s(2), \cdots, s(l_i))$ in the trellis of H_j, we determine the log likelihood, $l(\mathbf{y}_i, \mathbf{s}, H_j)$, that the observation $\mathbf{y}_i$ is generated via the state sequence, $\mathbf{s}$. Hence,

$$l(\mathbf{y}_i, \mathbf{s}, H_j) = \log \Pi_j(s(1)) + \sum_{t=1}^{l_i - 1} \log A_j(s(t), s(t+1)) + \sum_{t=1}^{l_i} \log B_j(s(t), \mathbf{y}_i(t)). \tag{1}$$

Here, $A_j(m, n)$ is the (m,n) element of the matrix, A_j. Similarly, $B_j(m, k)$ is the (m, k) element of matrix, B_j, and $\Pi_j(m)$ is the mth component of the vector, Π_j.

Next, we maximize the log likelihood over all state sequences in the trellis of HMM, H_j, and determine

$$d_j(\mathbf{y}_i) = \max_{\mathbf{s} \in \mathcal{S}_{l_i}(H_j)} l(\mathbf{y}_i, \mathbf{s}, H_j). \tag{2}$$

Here, $\mathcal{S}_l(H_j)$ is the set of all state sequences of length l in the trellis of HMM, H_j. The quantity, $d_j(\mathbf{y}_i)$ thus represents the log likelihood of the state sequence in model H_j, that most likely generated $\mathbf{y}_i$. Interpreting $d_j(\cdot)$ as the discriminant for class j, we adopt the traditional discriminant-based classification approach to define the classifier operation as :

$$C(\mathbf{y}_i) = arg \max_j d_j(\mathbf{y}_i). \tag{3}$$

We refer to this definition as the "best path" discriminant [1]. This classification system can be viewed as a competition between paths. The observation is ultimately labeled by the class index of the HMM to which the winning path belongs. One advantage of the "best path" discriminant classifier is that the search for the most likely path (choosing a state sequence, $\mathbf{s}$, that maximizes (2)) can be reduced to a sequential optimization problem that can be solved via an efficient dynamic programming algorithm (Viterbi search).

[1] Our design method can be easily modified to the case where the discriminant is obtained by appropriate averaging of the likelihood over all paths in the class model.

2.1 HMM classifier design

The problem of HMM classifier design can be stated as the joint optimization of the HMM parameters, $\{\Lambda_j\}$, to minimize the empirical probability of misclassification measured over the training set,

$$\min_{\{\Lambda_j\}} P_e = 1 - \frac{1}{N} \sum_{i=1}^{N} \delta(C(\mathbf{y}_i), c_i) \tag{4}$$

where δ is the error indication function: $\delta(u, v) = 1$ if $u = v$ and 0 otherwise.

The most important difficulty in this optimization is that the cost, P_e, is a piecewise constant function of the optimization variables. As a result, we cannot use traditional gradient descent based optimization methods - the gradients are zero almost everywhere. One approach [7] to circumvent this difficulty is to replace the piecewise cost function by a smooth approximation to it. While the modified cost function is amenable to descent-based optimization, in practice, there are numerous shallow local minima on the complex cost surface that can easily trap optimization methods based on simple descent. In the next section, we present a novel approach based on deterministic annealing to simultaneously tackle the piecewise nature of the cost function and the problem of shallow local minima traps.

3 Deterministic Annealing approach

We take as our starting point, the deterministic annealing approach to clustering, vector quantization [14] and related optimization problems [13] and its extension to structurally-constrained clustering problems [8]. The extended method can handle problems involving structural constraints on the clustering rule *e.g.* tree structured vector quantization, pattern classifiers based on parametric discriminant functions etc. We have recently applied the extended DA method successfully to the design of standard pattern classifiers [8], regression functions [9, 12, 11] and source coding systems [10]. The work presented in this paper represents an important extension of the method to handle time sequences that are modeled by HMMs.

We cast the optimization problem within a probabilistic framework and maintain that, during design, it is useful to consider a randomized HMM classifier system. In the randomized classifier, given an observation, a winning state sequence is randomly chosen from among all state sequences in all the HMMs. This (random) choice of the winning state sequence is based on a probability distribution - we replace the best-path discriminant rule which associates a pattern to a unique winning state sequence by a randomized best-path discriminant rule that associates each pattern, $\mathbf{y}_i$, to every state sequence, $\mathbf{s}$, in the trellis of every model, H_j, with a proba-

›ility, $P(\mathbf{y}_i, \mathbf{s}, H_j)$. Naturally, these probabilities are normalized such that $\sum_j \sum_{\mathbf{s} \in \mathcal{S}_{l_i}(H_j)} P(\mathbf{y}_i, \mathbf{s}, H_j) = 1$.

The probabilities, $P(\mathbf{y}_i, \mathbf{s}, H_j)$, are obtained in a systematic manner: We irst note that the non-random best-path discriminant rule may be expressed is minimization over $\mathbf{s}_i \in \bigcup_j \mathcal{S}_{l_i}(H_j)$ of the cost function,

$$D = \frac{1}{N} \sum_i l(\mathbf{y}_i, \mathbf{s}_i, H_j). \tag{5}$$

After randomness is introduced, this cost function is replaced by the expected :ost,

$$< D >= \frac{1}{N} \sum_i \sum_j \sum_{\mathbf{s} \in \mathcal{S}_{l_i}(H_j)} P(\mathbf{y}_i, \mathbf{s}, H_j) l(\mathbf{y}_i, \mathbf{s}, H_j), \tag{6}$$

which is minimized, while simultaneously enforcing a level of randomness though a constraint on the Shannon entropy,

$$H = -\frac{1}{N} \sum_i \sum_j \sum_{\mathbf{s} \in \mathcal{S}_{l_i}(H_j)} P(\mathbf{y}_i, \mathbf{s}, H_j) \log P(\mathbf{y}_i, \mathbf{s}, H_j). \tag{7}$$

In particular, we optimize $< D >$ subject to $H = \hat{H}$. The probability distribution obtained via this constrained optimization problem is the Gibbs distribution,

$$P(\mathbf{y}_i, \mathbf{s}, H_j) = \frac{e^{\gamma l(\mathbf{y}_i, \mathbf{s}, H_j)}}{\sum_{j'} \sum_{\mathbf{s}' \in \mathcal{S}_{l_i}(H_{j'})} e^{\gamma l(\mathbf{y}_i, \mathbf{s}', H_{j'})}}. \tag{8}$$

The value of Shannon entropy, $\hat{H}$, corresponding to this Gibbs distribution is determined by the positive *scale parameter*, γ. This parameter also controls the "randomness" of the distribution. For $\gamma = 0$, the distribution over paths is uniform. For finite, positive values of γ, the Gibbs distribution indicates that we assign higher probabilities of winning to state sequences with higher log likelihoods. In the limiting case of $\gamma \to \infty$, the random classification rule reverts to the non-random "best path" classifier, which assigns a non-zero probability of winning only to the path with the highest log likelihood as in (2).

The random classifier's expected rate of misclassification (over the training set) can be calculated as

$$< P_e >= 1 - \frac{1}{N} \sum_{i=1}^{N} \sum_{\mathbf{s} \in \mathcal{S}_{l_i}(H_{c_i})} P(\mathbf{y}_i, \mathbf{s}, H_j) \tag{9}$$

Next, we pose the problem of optimizing this random HMM classifier (choosing $\{\Lambda_j\}$ and γ) to minimize the expected mis-classification probability of (9). However, simply minimizing (9) over all Gibbs distributions chooses one that is non-random ($\gamma \rightarrow \infty$). While such a non-random, best-path classifier is the eventual goal of this design method, we wish to enforce the "non-randomness" gradually during the optimization, to avoid shallow local minima traps.

As such, we follow the philosophy underlying the deterministic annealing approach and pose the problem of minimizing $< P_e >$ while maintaining a level of randomness in the classifier through a constraint on the entropy, $H = \hat{H}$. This constrained optimization problem is equivalently, the minimization of the unconstrained Lagrangian cost function,

$$\min_{\{\Lambda_j\},\gamma} L \equiv < P_e > -TH, \tag{10}$$

where T is the Lagrange parameter that we refer to as the "temperature" because of an interesting analogy in statistical physics.

3.1 Analogy to statistical physics

The Lagrangian minimization of (10) reminds us of the definition of thermal equilibrium in statistical physics. The quantity, L, is analogous to the Helmholtz free energy of a thermodynamic system with average energy $< P_e >$, entropy over energy states, H and temperature, T. This free energy is the quantity that is minimized when this thermodynamic system is at thermal equilibrium at temperature, T.

From the optimization viewpoint, we are particularly interested in thermal equilibrium at $T = 0$ which corresponds to direct minimization of $< P_e >$, our ultimate objective. The analogy to physical systems suggests that to minimize $< P_e >$, it is useful to implement an annealing process, that is, gradually lower the temperature while maintaining the system at thermal equilibrium. We start with a very high value of T, where the sole objective is entropy maximization, which is achievable by the uniform distribution. Reducing T gradually from this high value, we repeat the process of minimizing L until $T = 0$, where the sole objective is optimizing $\{\Lambda_j\}$ and γ to minimize P_e.

After this annealing process, we also include as a final step, a "quenching" mechanism - we optimize $\{\Lambda_j\}$ to minimize P_e, while increasing γ from its optimal value at $T = 0$, in gradual steps, to a very high value. When γ is sufficiently high, the classifier reduces to the non-random "best-path" classifier.

The annealing process yields a sequence of solutions at decreasing levels of entropy and P_e leading to the "best-path" classifier in the limit. The DA method is not a stochastic method like simulated annealing, but instead based

on the optimization of the deterministically computed expectation, L, at each temperature. This minimization is achieved by a series of gradient descent steps with the following expressions for the gradients :

$$\frac{\partial L}{\partial \Lambda_j} = \frac{1}{N} \sum_i \sum_{\mathbf{s} \in \mathcal{S}_{l_i}(H_j)} L(\mathbf{y}_i, \mathbf{s}, H_j) P(\mathbf{y}_i, \mathbf{s}, H_j) \{ \frac{\partial l(\mathbf{y}_i, \mathbf{s}, H_j)}{\partial \Lambda_j} - < \frac{\partial l(\mathbf{y}_i, \mathbf{s}, H_j)}{\partial \Lambda_j} >_j \}$$

and

$$\frac{\partial L}{\partial \gamma} = \frac{1}{N} \sum_i \sum_j \sum_{\mathbf{s} \in \mathcal{S}_{l_i}(H_j)} L(\mathbf{y}_i, \mathbf{s}, H_j) P(\mathbf{y}_i, \mathbf{s}, H_j) \{ l(\mathbf{y}_i, \mathbf{s}, H_j) - < l(\mathbf{y}_i, \mathbf{s}, H_j) > \}$$

Here, $L(\mathbf{y}_i, \mathbf{s}, H_j) = T\gamma l(\mathbf{y}_i, \mathbf{s}, H_j) - \delta(j, c_i)$. The operation, $< f(\cdot) >_j$, represents an expectation of the (state-sequence dependent) $f(\cdot)$ function over the state sequences in the trellis of HMM, H_j. Hence,

$$< \frac{\partial l(\mathbf{y}_i, \mathbf{s}, H_j)}{\partial \Lambda_j} >_j = \sum_{\mathbf{s} \in \mathcal{S}_{l_i}(H_j)} P(\mathbf{y}_i, \mathbf{s}, H_j) \frac{\partial l(\mathbf{y}_i, \mathbf{s}, H_j)}{\partial \Lambda_j} \tag{11}$$

Similarly, $< f(\cdot) >$ represents the expectation of the $f(\cdot)$ function over all state sequences in the trellises of all the HMMs. Hence,

$$< l(\mathbf{y}_i, \mathbf{s}, H_j) > = \sum_j \sum_{\mathbf{s} \in \mathcal{S}_{l_i}(H_j)} P(\mathbf{y}_i, \mathbf{s}, H_j) l(\mathbf{y}_i, \mathbf{s}, H_j). \tag{12}$$

An important aspect of the proposed method is the discovery of an efficient forward-backward algorithm to determine these gradient parameters. Note that the summations in the gradient expressions are over all state sequences in the trellis of HMMs. The number of paths depends exponentially on the number of states in the HMM. However, these summations can be efficiently computed via a forward-backward algorithm which reduces the number of computations substantially (proportional to square of the number of states in the HMM) thus cutting down on computational complexity and memory requirements. The complexity of the DA method scales similarly to the maximum likelihood method with respect to the number of states and training vectors.

4 Experimental Results

We have performed preliminary simulations to determine the usefulness of our new design method. We experimented on designing simple (2,3 and 4

class) classifier systems for eight different data sets of 2000 vectors each. Allowing three to six states in each Markov model, we designed HMM classifier systems using the maximum likelihood and deterministic annealing methods. We observe that the proposed DA approach improved the classification performance consistently and considerably. Over the experiment's data sets, the rate of misclassification was reduced by factors of 1.2 to 3. Table 1 details the results.

We are currently investigating the effectiveness of the design method on real-world speech data to demonstrate its advantages for the speech recognition problem.

Dataset	1	2	3	4
No. of Classes	2	2	2	3
P_e (ML)	17.4%	31.6%	26.5%	28.7%
P_e (DA)	6.5%	21.7%	18.7%	20.9%
Dataset	5	6	7	8
No. of Classes	3	3	3	4
P_e (ML)	27.0 %	32.5%	24.9 %	42.3 %
P_e (DA)	21.0%	27.3%	17.4%	31.7%

Table 1: A comparison of the mis-classification rates obtained for HMM classifiers designed from eight classified training sets of 2000 patterns each. Each set consists of data from 2,3 or 4 classes. ML represents a Max. likelihood design algorithm and DA represents the deterministic annealing algorithm.

5 Conclusion

In this paper we propose a novel training method for HMM classifier systems that jointly optimizes all the models to minimize the true cost, namely, the rate of mis-classification. At the cost of moderate increase in complexity, considerable improvements in recognition rates are obtained.

References

[1] L.R. Bahl, P.F. Brown, P.V. DeSouza, R.L. Mercer, "A new algorithm for the estimation of hidden Markov model parameters", Proc. ICASSP-88, pp. 493-496.

[2] L.R. Bahl, P.F. Brown, P.V. DeSouza, R.L. Mercer, "Maximum Mutual Information estimation of hidden Markov model parameters", Proc. ICASSP-86, pp. 49-52, Tokyo, Japan.

[3] J.K. Baker, "Stochastic modeling for automatic speech recognition", In D.R.Reddy, ed., Speech Recognition, New York: Academic Press, pp. 521-542, 1975.

[4] H. Boulard, Y. Konig, N. Morgan, "REMAP: Recursive Estimation and Maximization of A Posteriori Probabilities - Application to Transition-Based Connectionist Speech Recognition", ICSI technical Report TR94-064, Internat, Computer Science Inst. CA.

[5] H. Watanabe, S. Katagiri, "HMM speech recognizer based on discriminative metric design", Proc. of ICASSP, 1997, vol. 4, pp. 3237-3240.

[6] F. Jelinek, "Continuous speech recognition by statistical methods", Proceedings of the IEEE, vol. 64, pp. 532-556, Apr. 1972.

[7] B.H. Juang, and S. Katagiri, "Discriminative learning for minimum error classification", IEEE Trans. on Sig. Proc., vol. 40, pp 3043-3054, 1992.

[8] D. Miller, A. Rao, K. Rose, A. Gersho, "A Global Optimization Technique for Statistical Classifier Design", IEEE Transactions on Sig. Proc., vol. 44, pp 3108-3122, 1996.

[9] A. Rao, D. Miller, K. Rose, A. Gersho, " An Information-theoretic Approach for Statistical Regression with Model Growth by Bifurcations", Computing Science and Statistics, Proc. of the 27th Symposium on the Interface, 1995, pp 220 - 224.

10] A. Rao, D. Miller, K. Rose, A. Gersho, " A Generalized VQ Method for Combined Compression and Estimation", Proc. of the IEEE ICASSP, 1996, pp. 2032 - 2035.

11] A. Rao, D. Miller, K. Rose, A. Gersho, " A Deterministic Annealing Approach for Parsimonious Design of Piecewise Regression Models ", submitted for publication.

12] A. Rao, D. Miller, K. Rose, A. Gersho, "Mixture of Experts Regression Modeling by Deterministic Annealing", accepted for publication in the IEEE Transactions on Signal Processing, Nov. 1997.

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

14] K. Rose, E. Gurewitz, G.C. Fox, " Vector quantization by deterministic annealing ", IEEE Trans. on Inform. Theory, vol. 38, pp 1249-1258, 1992.